Diagnosis auto-generation based on Pathological Images with Deep Learning

Jianqiao Tian
Department of Biomedical Engineering
University of Florida

Yun Liang
Department of Biomedical Engineering
University of Florida

***Abstract*—Medical image diagnosis, especially histopathology images, is usually conducted by pathologists' careful review of histopathology images. This is usually accomplished by dictating the data into the text. However, this conventional methods are error-prone and time-intensive, due to the leak of well-experienced professionals and variability of various disease. To address this challenge, we propose one approach combining image extraction and report generation. First, we ultize a pre-trained ResNet to extract image features from the medical image patches with the size of 512*512. The extracted image feature is arranged as a vector with size of 512, 1024, 2048, and 4096. Secondly, this vector, along with training captions, will be feed into LSTM for report generation. We demonstrate the effectiveness of the proposed methods on one bladder histopathological image datasets with our experiments. A brief discussion regarding current drawback and future direction is also provided.**

I. Introduction

Medical images, such as Computed Tomography (CT), Magnetic resonance imaging (MRI), Positron emission tomography (PET), and histopathological images, are all commonly used images in hospitals and clinics for disease diagnosis. Specific medical fields usually conduct the reading and interpretation of the above medical images to make a diagnosis for these images. When analyzing these images, it usually involves a large number of images and many more high-level image features to make an accurate and keen diagnosis. Take histopathological bladder images used in this work as an example, and diagnosis is generally made based on observations from perspectives such as cell distribution, cell variation, nuclear condition, cell cycles, etc. A final diagnosis is the result of these varying observations. However, this conventional work-flow shows inefficiency when considering how labor-intense and time consuming this conventional work-flow could be. Also, due to the leak of well-experienced professionals and variability of various disease, making a medical images diagnosis is challenging, and often error-prone, for less-experienced professionals.

In this study, we specifically investigate bladder biopsy histopathological images. These biopsy slices are taken from bladder tissue and used for bladder cancer screening. The observation in this specific task involves the degree of pleomorphism, the sign of crowding nuclei, the polarity of membranes, and the appearance of nucleoli. A final evaluation is made to reflect varying degrees of cancer-like, i.e., "normal", "low grade", "high grade", and "insufficient information". The ultimate goal of this study is to utilize deep learning techniques to mimic this work-flow, i.e. when given a histopathological biopsy image, and the proposed machine should be able to return a caption to reflect these observations accurately.

Deep learning, which has been commonly applied to natural image classification and segmentation tasks, is continuously reported to out-performer many other machine learning approaches, as well as some human counterpart for some specific tasks. Thus, we are motivated to use deep learning framework for our computer-aided medical report generation. One of the most state-of-the-art methods is published in 2017 [4], with the combination of standard Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) with an attention mechanism. Another work on medical report generation has moved beyond sentences and attempted the difficult task of generating long and topic-coherent reports to describe visual content which is much more difficult. To further extend the medical report generation, we propose one medical report auto-generation framework via combining of ResNet with LSTM.

The rest of this report is arranged as follows: in Sec.2, we provide some related works, from which we obtain our inspirations. Followed by a detailed introduction regarding the methodology we use in Sec.3. The utilized Database and specific experiment configuration are introduced in Sec.4 and five respectively. Last but not least, a brief discussion based on the returned result is given in Sec.6, to support our conclusion made in Sec.7.

II. Related Work

Medical image diagnosis auto-generation problem can be viewed as a special case of image captioning. Both of them contains two major parts: image feature extraction and subsequent caption generation. In this work, we employ CNN to achieve images feature extraction, and LSTM to generate a diagnosis report in the form of natural language.

Image captioning Image captioning is a combination of natural language processing and computer vision. It aims at automatically generating text descriptions for given images. Generally speaking, image captioning can be broken into two parts: image feature extraction, and natural language sequence generator. These two parts are often referred to as encoder and decoder in image captioning problems. Most recent image captioning models are based on a CNN-RNN framework[10]

Impression: No acute cardiopulmonary abnormality.

Findings: There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. There is no evidence of pneumothorax. Degenerative changes of the thoracic spine.

MTI Tags: degenerative change

Fig. 1: An example chest x-ray report. In the impression section, the radiologist provides a diagnosis. The findings section lists the radiology observations regarding each area of the body examined in the imaging study. The tags section lists the keywords which represent the critical information in the findings. These keywords are identified using the Medical Text Indexer(MTI).

[2] [5] [11]. CNN is considered as the go-to method for image-related problems since it can dramatically reduce the number of parameters that need to be trained when it comes to images. And RNN, thanks to its ability to handle sequential information, is the strong candidate for generating natural languages sequence. LSTM is more updated version for RNN with stronger capability of generating natural languages.

Besides, the attention mechanism has been widely applied in both CNN and RNN frameworks. [10] has recently proposed a spatial-visual attention mechanism over image features extracted from intermediate layers of the CNN. [12] propose a semantic attention mechanism over tags of given images.

Medical Image Diagnosis Before medical diagnosis, a previous problem called medical image labeling has been a well-studied problem. These labeling tasks, either fully-structured or semi-structured, are more like an extension of image classification [6]. In traditional image classification problems, one image is usually assigned with the only label. While image labeling can attach one image with multiple different names in the different genre.

Image captioning takes this generated labels into a natural language level. Therefore, two major parts are involved in the overall work-flow: extracting image information and deciding the whole sequence based on this extracted image information. One sample for check X-Ray diagnosis report is shown in Fig 1.

The practical importance of this problem is somehow significant. Conventionally, the quality of these medical diagnosis varies from physicians to physicians. Also, some physician could be much more specialized in some area/disease, but less experienced in others. Such a limitation can be expected to be significantly assisted with deep learning which can automatically generate a medical diagnosis. Even though deep learning techniques are not the same as human performance on a broad base, they have advantages that the average human cannot compete either. For example, deep learning can take a

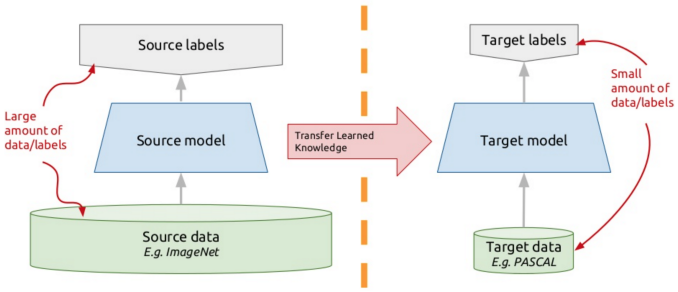

Fig. 2: The fundamental philosophy of transfer learning.

considerable amount of data into consideration during training. Besides, deep learning methods can gather multiple knowledge from various datasets in a shorter period, which is another advantage that human competitors cannot keep pace with. Due to the unbalanced distribution of health care resource, reliable diagnosis generation methods will be much more valuable in computer-aided diagnosis.

Transfer Learning Transfer learning referrers to a situation where training database and testing databases have different fundamental properties. In this study, transfer learning exists in the image feature extraction part. The images feature is extracted by ResNet, pre-trained on ImageNet. ImageNet is vast image pool of natural images, with corresponding labels. This database is originally built for natural image classification challenge. Even though our problem is neither a classification problem nor a natural images problem, this pre-trained network can still work in our favor. This is because we assume the lower layers of ResNet captures low-level images features, such as short edges along with different directions, and these lower level features will not change much for both natural images and medical images. Especially, for histopathological slices, which looks rather similar to natural images, compared with CT, or MRI, the low level images features should be sufficiently universal between histopathological images and natural images. The basic philosophy of transfer learning is shown in Fig 2.

The most significant impact for transfer learning is that transfer learning enable us to deal with specific problems with very minimal number of required data. In our experiment, we have 2364 slice images for training. ResNet-152, which has 152 layers, has 60 million parameters that ought to be trained. Without a large database, such as ImageNet, it would be impossible to train such a large network on only 2364 images. Transfer learning is critically important and useful for medical image problems, since medical images are less available than natural images.

III. Methodology

The proposed methodology divided into two parts: image feature extraction and text generation. The whole framework is shown in Fig 3.

Image Preprocessing For each autopsy images, after the tissue dyed, these tissues are placed under the con-focal

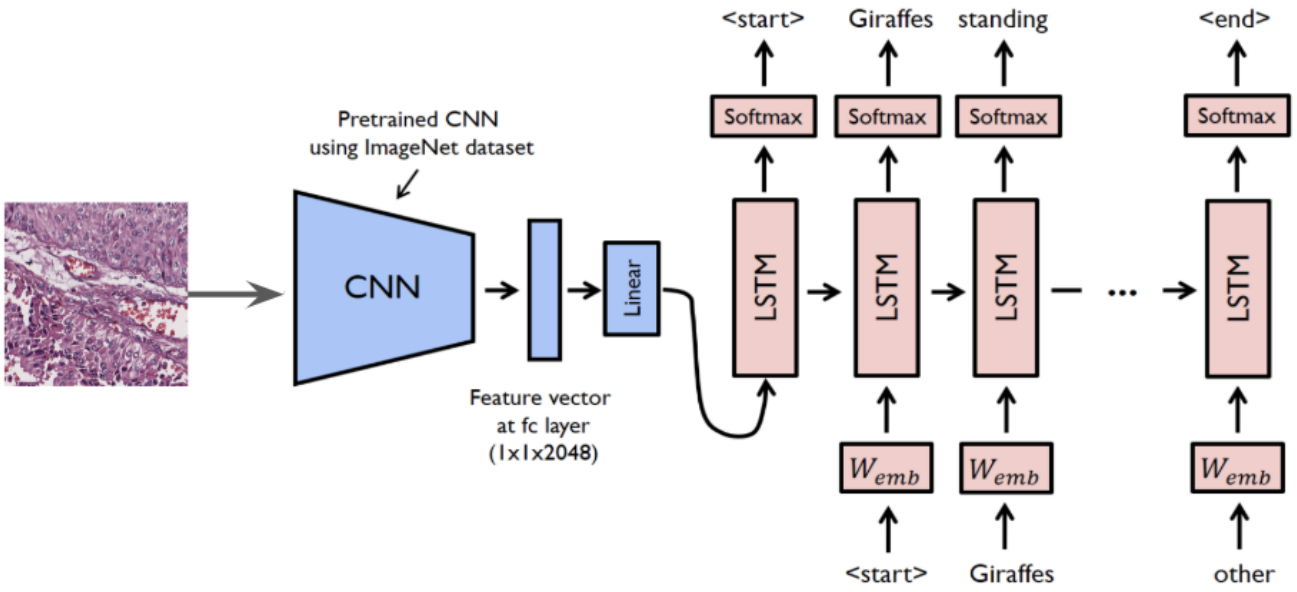

Fig. 3: The proposed framework is shown in the figure. We will first extract image features, and then apply the extracted feature to feed into LSTM.

microscope to take these histopathological images. Each whole slice images are divided into multiple patches, with size as 512*512 pixels. Patches that possess tissue less than 70 % have been discarded. Besides these, we did not employ any other pre-processing technique to these histopathological images. This can be taken as an illustration of our method's generality since it does not require special pre-processing techniques.

Text Generation Text generation is based on the dependency between each work within a sequence. LSTM captures this dependency. And this dependency is used to give Sequential dependency between phrases/words, such as given the phrase "crowded to", the next word should be one of the following: "mild, moderate, high" degree. However, the decision of which word the sequence should adopt is decided based on the image feature extracted from input images.

ResNet A Convolutional Neural Network (CNN, or ConvNet) is a special kind of multi-layer neural networks, designed to recognize visual patterns directly from pixel images with minimal preprocessing, which is the current state-of-the-art approach in image classification and segmentation. The detailed structure is shown in Fig 4. There are three versions of ResNet, and we use pretrained ResNet-152 as our feature extraction choice.

LeNet[1], Alexnet[3], ZFNet[13], ResNet[9], vggnet [8] are many current widely-used convolution neural network. In our approach, we use ResNet as our image extraction network. The detailed ResNet structure is shown in Fig 5

The ResNet architecture is often recognized as the state-of-the-art in image classification tasks. The fundamental building unit of the ResNet architecture is the ResNet block, as shown in the below Fig 6.

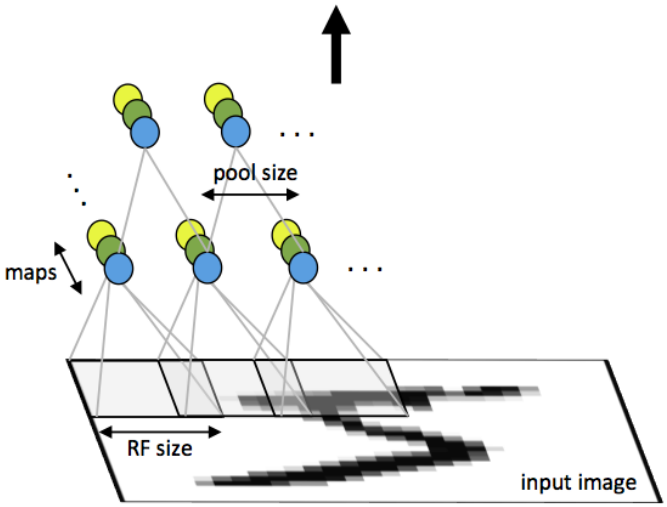

Fig. 4: First layer of a convolution neural network with pooling. Units of the same color have tied weights and units of different color represent different filter maps.

Recurrent Neural Network(RNN) Sequential information is used in RNN and related LSTM. In traditional neural networks, we assume that all inputs (and outputs) are independent of each other. However, this is not reliable in tasks that focus on sequential problems. The relationship among different words is of great help to predict the next sentence. RNN utilizes the relationship among multiple words, and LSTM includes more updated gates including the forget gate to use the sequence information better. Fig. 7 shows what a

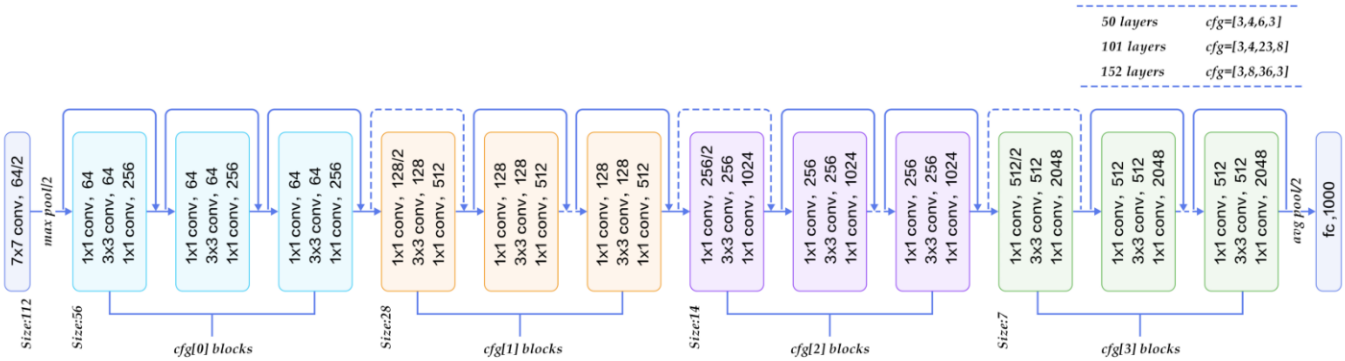

Fig. 5: At the ILSVRC 2015, the so-called Residual Neural Network (ResNet) by Kaiming He is introduced anovel architecture with skip connections and features heavy batch normalization.

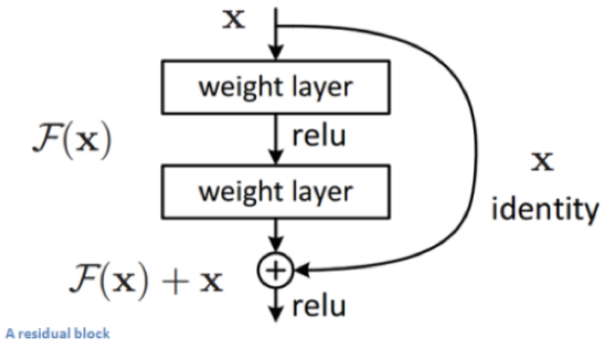

Fig. 6: Resnet Block.

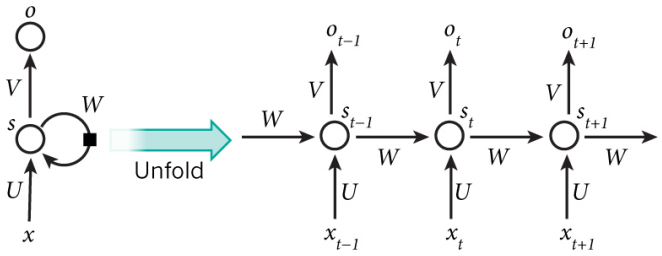

Fig. 7: A recurrent neural network and the unfolding in time of the computation involved in its forward computation.

typical RNN looks like.

Image Feature Extraction The model uses pre-trained ResNet-152 as our Convolution Neural Network for feature extraction. As in the original ResNet outputs 1000 labels, we reapply the last layer with the size of 1*1024 vectors. We use 512,1024,2048 and 4096 as our output feature, and we perform multiple experiments on utilizing the performance of various output features.

IV. Experimental Dataset

To evaluate the proposed framework, we conduct experiments on one bladder cancer database. We use image patches for the whole experiment. The whole dataset includes 2364 training images and 1888 testing images with the size of 512*512. For each image, there are five diagnosis reports based on the diagnosis of five separate pathologists. Besides, all the patches are divided into four classes: normal, low grade, high grade and insufficient information(unsure). For data processing, the input image is resized to 224224. We subtract the RGB mean from each image and augment the training data through clip, mirror and rotation operations. Sample data are shown in Fig 8.

The sample annotation is shown below.

Sample Annotation ”N2_07_2”: {”caption”: [”Mild pleomorphism is present. There are no signs of crowding in the nuclei. Polarity along the basement membrane is negligibly lost. Mitosis is,rare. The nuclei have inconspicuous nucleoli. Normal.”, ”Nuclear,features show mild pleomorphism. Pictured nuclei exhibit normal crowding.,Polarity is not lost. Mitosis are exceedingly rare and limited only to the,basal layer of urothelium. The nuclei have inconspicuous nucleoli.,Normal.”, ”Nuclear features show mild pleomorphism. There are no,signs of crowding in the nuclei. The nuclei retain a normal polarity. Mitosis,are exceedingly rare and limited only to the basal layer of urothelium. The,nuclei have inconspicuous nucleoli. Normal.”, ”Mild pleomorphism,and cytologic atypia is present. There are no signs of crowding in the,nuclei. Polarity along the basement membrane is negligibly lost. Mitosis,appears to be rare. The nuclei have inconspicuous nucleoli. Normal.”,,”Mild pleomorphism and cytologic atypia is present. There are no signs,of crowding in the nuclei. Polarity of nuclei is negligibly lost. Mitosis is,rare. Nucleoli is absent to inconspicuous. Normal.”],”label”:0 }

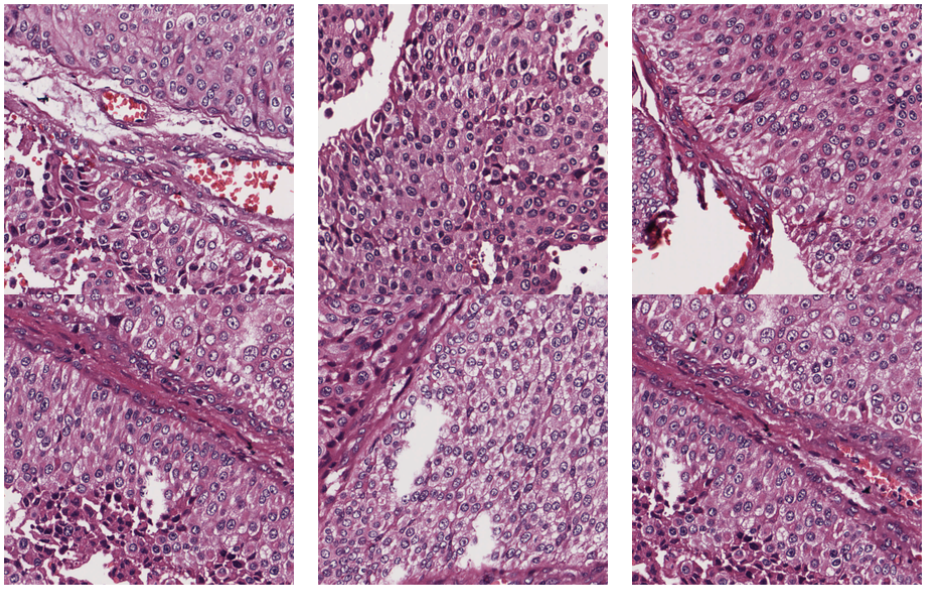

Fig. 8: All the image patches are with the same size of 512*512, extracted from 200 whole slide images.

V. Experiment

Experiment setup We perform our experiments with one framework that can be viewed as a combination of two separate networks. One part is ResNet-152 which is pretrained on ImageNet as our image feature extraction network. The later text generation network is based on one dimension of LSTM. We utilize nltk package for LSTM embedding and text generation. To compare the different extracted vector length, we have applied 512,1024,2048 and 4096 as image extracted features size and use the same hyperparameters and one layer of LSTM with the same embedding dimension of 256. All the experiments are trained for 30 epoches, and the hyperparameters for Resnet-152 are set the same as that pretrained in imagenet.

Evaluation Matrix BLEU (Bilingual Evaluation Understudy)[7] is a measurement of the differences between an automatic translation and one or more human-created reference translations of the same source sentence. In our experiment, we use the BLEU score as our evaluation method for comparing with the original caption and the generated caption. The detailed formula for BLEU score calculation is as below:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram} Count_{matches}(ngram)}{\sum_{S \in C} \sum_{ngram} Count(ngram)} \quad (1)$$

where n-gram precision p_n by summing the n-gram matches for each hypothesis sentence S in the test corpus C. A higher BLEU score means higher similarity of two captions, while lower BLEU score demonstrates lower relationship. As each patch has five separate captions, we choose one diagnosis randomly from the five diagnoses and compare the BLEU score between our generated caption and original caption.

VI. Result and discussion

Our experiment result, with image feature length as 512, 1024, 2048, and 4096, is shown in Figure 9.

The histogram result is shown in Figure 10.

The box plot result is shown in Figure 11.

The generated caption is shown as below: We choose image 115872_012 as our sample image. The image is shown in Fig 12: The caption generated by 512 vector is: **moderate pleomorphism and cytologic atypia is present . moderate nuclear crowding is seen . polarity is not completely lost**, and the BLEU score is 0.7125.

For 1024 length vectore, the generated caption is: **insufficient information . insufficient information . insufficient information . insufficient information . insufficient information . insufficient information . insufficient** with a BLEU score of 0.097.

The caption generated by 2048 vectore length is: **nuclear features show moderate pleomorphism . there is a moderate degree of crowding . polarity is not lost .** with a BLEU score of 0.45.

Finally, the caption generated by 4096 vectore length is: **nuclear features show moderate pleomorphism . there is a mild degree of crowding . polarity is not completely lost** with a BLEU score of 0.45.

It can be seen that in this specific caption, using 512 vectors provides the best result with the highest BLEU score; however, using the 2048 and 4096 provides almost the same results. Using 1024 includes kind of a nonsense result, and it seems that this does not work in this specific caption. Besides, from the caption histogram comparison, it appears that only using 512 vectors has a higher result. This result matches the sample image annotation result.

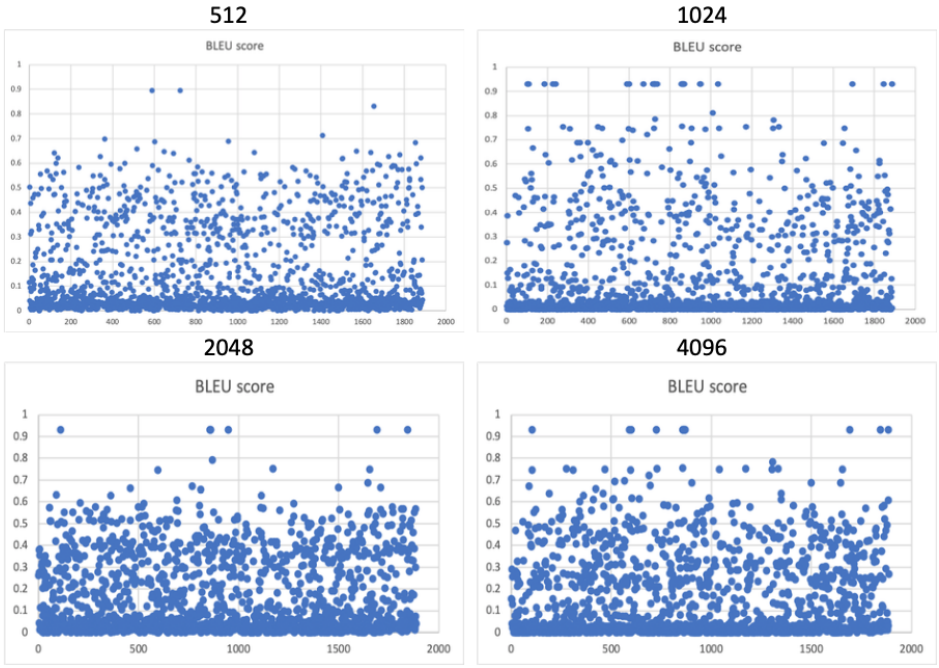

Fig. 9: The overall result are shown in the figure. It demonstartes the overall result for 512,1024,2048 and 4096 features generated.

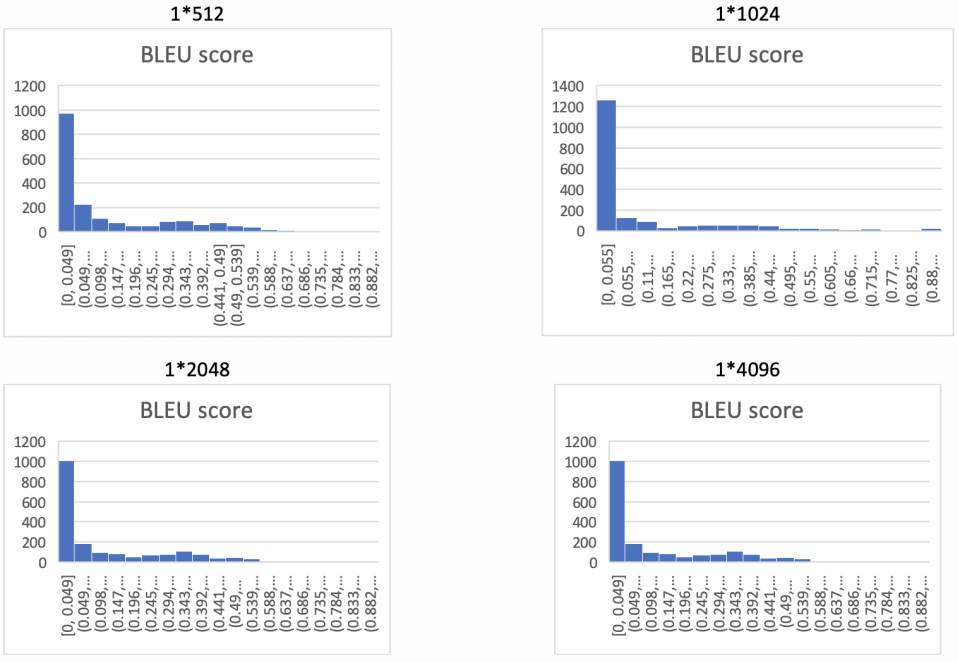

Fig. 10: The histogram results of four experiments.

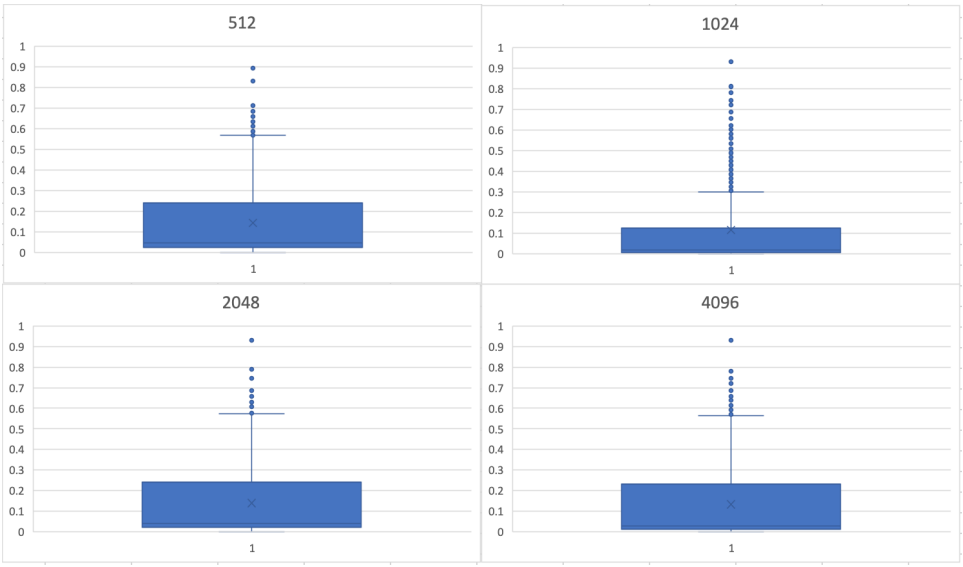

Fig. 11: The box plot results of four experiments.

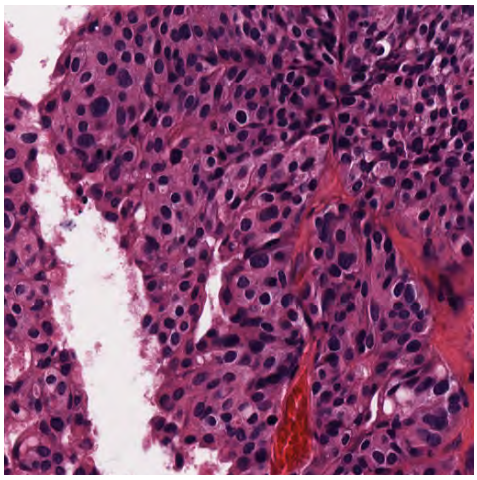

Fig. 12: Sample image.

From which we can see that 1:) our proposed method do have the ability to capture the fundamental rule from given captions informs natural language sequences. And this proposed method could achieve the BLEU score as high as 0.94. However, the majority of this return results concentrated at a lower BLEU score ranges.

This pair of generated caption and truth is a very general demonstration of the overall results. As we observe, most of the captions that have low BLEU scores are because the latter part of the caption is totally lost, rather than wrong.

This observation indicates two important findings: First, the image feature has been successfully extracted. If the image features are not extracted successfully, then the returned caption should contain a higher degree of errors, instead of missing parts. This is not the case as we observed. The low BLEU score is the result of an incomplete caption.

Second, the LSTM layer used in our current experiment is not long enough. Since the later part of the caption is missing, we assume a less-than-optimal structure of RNN causes this. Also, during our experiment, we used only one caption instead of all five captions. It is suspected that if all five captions for one image are fully utilized, the RNN can learn this repeated pattern better, and therefore, generate a full-length caption. This point is preserved as a future work direction.

VII. Conclusion

Our pilot study has shown that our proposed method, i.e., combining ResNet and LSTM, has a strong potential to accurately generate image diagnosis for bladder histopathology images for bladder cancer diagnosis. Yet, there still is a wide space for us to improve our proposed work to achieve higher accuracy. As analyzed above, the reason of low accuracy is assumed to be the RNN's incapability of complete caption sequence. With a more dedicated design of RNN, such as adding attention mechanism would possibly improve the performance of our proposed method.

Author Contribution

Our project is divided equally. Both of us work on the algorithm implementation part, and for the four experiments, each of us perform two experiments. For project report, we seperate the report into two equal parts, and both of us work on the project report.

REFERENCES

[1] Al-Jawfi, R.: Handwriting arabic character recognition lenet using neural network. Int. Arab J. Inf. Technol. 6(3), 304–309 (2009)

[2] Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. arXiv preprint arXiv:1505.01809 (2015)

[3] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)

[4] Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)

[5] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)

[6] Kisilev, P., Walach, E., Barkan, E., Ophir, B., Alpert, S., Hashoul, S.Y.: From medical image to automatic medical report generation. IBM Journal of Research and Development 59(2/3), 2–1 (2015)

[7] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)

[8] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[9] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

[10] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)

[11] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057 (2015)

[12] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4651–4659 (2016)

[13] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)