Feature extraction and Cluster analysis of Pancreatic Pathological Image Based on Unsupervised Convolutional Neural Network

Konosuke Asano*, Naoaki Ono†, Chika Iwamoto‡§, Kenoki Ohuchida¶ Koji Shindo¶ Shigehiko Kanaya‖
*Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
†Data Science Center, Graduate School of Information Sciences, Nara Institute of Science and Technology, Nara, Japan
‡Department of Advanced Medical Initiatives, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan
§Center for Advanced Medical Innovation, Kyushu University, Fukuoka, Japan
¶Department of Surgery and Oncology, Graduate School of Medical Sciences, Kyushu University,Fukuoka, Japan
‖Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan
*asano.konosuke.ac0@is.naist.jp

***Abstract*—In recent years, computer-aided diagnosis based on machine learning, mainly based on Convolutional Neural Network (CNN) has been studied and developed rapidly. Those methods are not only helpful for classification, but also useful for feature extraction from given images, especially encoding image data into discrete representation helps us obtain new knowledge. Previous researches showed that CNN can by trainded for not only dection of cancers but also classification of gene expression subtypes. Although most of these studies are based on supervised learning that needs curated pathological knowledge, it is useful to extract characteristic features in the given images, using unsupervised machine learning in order to obtain new pathological findings. We applied cluster analysis using CNN which is trained based on adversarial training and maximization of mutual information and showed that it can classify those pathological images into discrete categories. Next, we applied our model for comparison of the two staining method in order to evaluate the degree of malignancy according to fibrosis and cell differentiation. The results showed that encoding of the histopathological image into discrete representations helps us to interpret tumor images.**

***Index Terms*—Convolutional Neural Network(CNN), Histopathology, Computer-Aided diagnosis(CAD)**

I. Introduction

Pathologists manually inspect a huge number of samples every day and make a diagnosis. Long hours of work will cause a decrease in concentration and increase the risk of misdiagnosis. Therefore, studies of computer-aided diagnosis based on machine learning have been studied intensively. Most studies [1] of medical image processing based on supervised learning based on heuristically designed models that need appropriate labels according to the task, the prediction from the model changes depending on how to label it. At the clinical site, it is desirable to identify tumors, but also to evaluate the types and stages of the tumor in order to decide appropriate treatment plan. Pancreatic cancer is known as one of the poorest prognosis cancer, due to the difficulty of early detection, the fast progression stages and the frequency of distant metastasis, etc. Therefore, estimation of medical features of the tumor cell, such as the proliferation rate, metastatic state, etc. will provide much advance to lead to better treatment. Tumor tissues are generally heterogeneous, composed of a mixture of various types of cells, it implies that the tumor futures should be recognized from the local variation of morphological features of the cells, such as the shapes and distribution of cell types in each region in order to help a doctor's diagnosis.

In preliminary experiments, we extracted phenotypic features from HE stained images with Autoencoder (AE) and Principle Component Analysis (PCA). When we analyzed small patches of images which are randomly cropped from the whole pictures, the features ranged widely and it was difficult to categorize them. We applied cluster analysis using Convolutional Neural Network (CNN) which is trained based on virtual adversarial training [2] and maximization of mutual information [3] and showed that it can classify those pathological images into discrete categories. Next, we applied our model for comparison of the two staining method in order to evaluate the degree of malignancy according to fibrosis and cell differentiation. The results showed that our model is able to obtain discrete representation related to these malignant indicators.

This work was supported by JSPS KAKENHI Grant Number 17H05297.

II. Material and Methods

We used pancreatic cancer pathological images of KPC mouse, a model mouse that follows a process similar to human pancreatic cancer. Generally, when visualizing sliced tissue specimens, pathologists use hematoxylin & eosin (HE) staining methods to observe nucleus and cytosols, and other immunological staining methods are additionally applied for diagnosis of several endocrine cancers.

In our study, we compared two different dying methods, HE staining which is one of the basic staining method and Massons trichrome (MT) staining of the contact slices of the pancreas tumors in KPC mouse. By HE staining, hematoxylin

stains the cell nucleus purple and eosin stains the fibers pink, which is used to visualize the structure of the cell tissue. On the other hand, MT staining highlights the cell nucleus reddish violet as well as HE, and collagen fibers, those are not stained by HE, in blue.

In our dataset, every MT-stained images are registered against the images of the closest slices with HE staining. Though the original images were quite high resolution (about 60K $\times$ 100K pixels), we randomly cropped small patches of 1024 $\times$ 1024 pixels and resized these into 224 $\times$ 224 to feed to the model. To crop small images from a large pathological image is widely used in the previous studies [1]. The number of training paths was 12000, and 3000 for the test evaluation. We used two clustering methods. One is based on Autoencoder and adversarial training, the other is based on self-augmented training and mutual information maximization.

A. Information Maximizing Self-Augmented Training (IMSAT)

We conducted cluster analysis using Information Maximizing Self-Augmented Training (IMSAT) [4]. Although IMSAT is an unsupervised learning, it can perform clustering with high accuracy. It is based on Self-Augmented Training and maximization of mutual information. When data is distorted by any affine transformation and perturbations, local representations in the latent space departs greatly from the predictions of original data points [5]. Using SAT, it can be close the predictions of perturbed data points to original data points. KL-divergence D_{KL} is used as the distance between original and distorted distribution. We choose Virtual Adversarial Training (VAT) [2] and typical data augmentation such as random contrast enhancement, rotation, flipping, and scaling as distortion in SAT. Adversarial perturbation r_{vadv} are generated in VAT. These disturb the prediction distribution greatly and can be computed from multiplying random normalized vector $\bar{d}$ by gradients $\nabla D_{\mathrm{KL}}[p(y|x), p(y|x+\bar{d})]$. Finally, loss functions of SAT are described as follows,

$$\mathcal{R}_{\mathrm{SAT}}(\theta; T) = \frac{1}{N}\sum_{x\in\mathcal{D}} \mathbb{E}_r[D_{\mathrm{KL}}[p(y|x,\theta), p(y|T(x),\theta)]] \quad (1)$$

where $T(x)$ is general distortion function.

In addition to this, penalty term of Regulized Information Maximization (RIM) [3] is applied for clustering. The RIM learns a probabilistic classfier $p_\theta(y|x)$ such that mutual information [6] between inputs and cluster assignments is maximized. Mutual information is a statistical dependency of two random variables. Finally, loss functions of IMSAT are described as follows,

$$\mathcal{R}_{\mathrm{SAT}} - \lambda[H(Y) - H(Y|X)], \quad (2)$$

where $H(Y)$ and $H(Y|X)$ are entropy and conditional entropy. The two entropy terms can be calculated as

$$H(Y) = h\left(\frac{1}{N}\sum_{i=1}^{N} p_\theta(y|x_i)\right), \quad (3)$$

$$H(Y|X) = \frac{1}{N}\sum_{i=1}^{N} h(p_\theta(y|x_i)), \quad (4)$$

$$h(p(y)) = -\sum_{y'} p(y')\log p(y') \quad (5)$$

Increasing the marginal entropy $H(Y)$ encourages the cluster sizes to be uniform, while decreasing the conditional entropy $H(Y|X)$ encourages unambiguous cluster assignments [7]. In our model, the method used in [4] was changed and the perturbation range ϵ was calculated using the centroid in the minibatch and the Euclidian distance between every sample.

B. Experiments

Inputs were preprocessed to be [0,1] range since we extracted feature using 50 layer deep residual networks (ResNet) [8] which was pre-trained with a subset of ImageNet. All parameters except final hidden layer in ResNet are frozen and additionally defined layers are fine-tuned with 50 epoch. Since it is an exploratory analysis of a pathological image, we trained models with the cluster size k as four conditions of 4, 8, 16 and 32.

III. Results & Discussion

In this section, we show the results of clustering and analysis of latent space using manifold learning.

A. Clustering

To avoid mixing of too large variety of features in the same cluster, we started dividing into more than four clusters $k = 4$. And when we increased the number of clusters ($k = 32$), some clusters vanished since gradients vanishing problem. Actually, it becomes difficult to interpret features of each cluster when it is too detail, we stopped dividing cluster up to 8. We concluded that the optimal cluster size in our dataset is less than 8. Therefore, we discuss the results of cluster size 8 in this section.

As (Fig. 1) shows, we can see trend of each cluster under the condition of $k = 8$. In cluster 0 and 1, these express similarities with other. These clusters are considered as hyalinization region and typical pancreatic cancer region. Moreover, we can see much stroma and less differentiation in these clusters. Hyalinization indicates a region where collagen fibers are fused and stained homogeneously with eosinophila. A hyalinization is one of the indicators how much cancer is progressing. In cluster 2 and 4, these are the edge of the tissue and sparse region. Although there is a subtle difference, we couldn't distinguish between cluster 2 and 4. In cluster 3, we can see normal glandular tissues which aren't cancer region. In cluster 5, large vessel walls are characteristic. In cluster 6, there are many edges of the tissues. The difference from the clusters 2,4 and 6 is that blank regions exist in clusters 2 and 4, while adipose tissues and inflamed tissues are seen in cluster 6 In cluster 7, indicates less stroma and progressed cell differentiation region. In this region, although the density of cancer cells is high, it is metastasis and invasion hardly occur since there are few interstitial tissues.

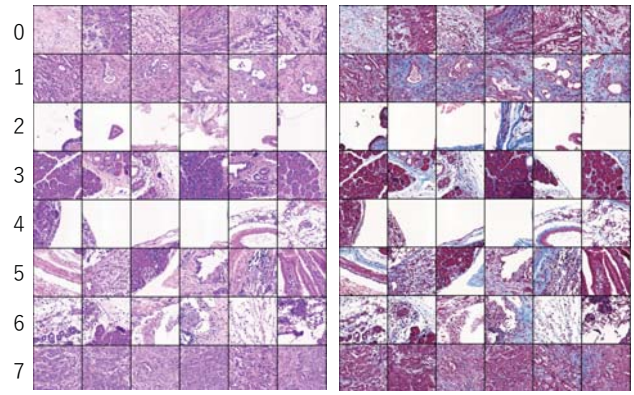

Fig. 1. Clustering results with IMSAT ($k = 8$). (Left) HE-stained images, (Right) MT-stained images.

B. Latent space

We visualized the latent space of the trained model by reducing dimensions using Isometeric Mapping (Isomap) [9] which is one of manifold learning. As shown in (Fig. 2), it can be seen that the characteristic cluster 3 and the other clusters can be clearly separated. x becomes smaller in latent space, we can see that images has a dense feature, and as the x becomes larger, the image have a sparse feature. It is also interesting that the cancer tissue and the normal glandular tissue are clearly separated in the latent space. As the results show, encoding the histopathological image into discrete representations helps us to interpret the pathology.

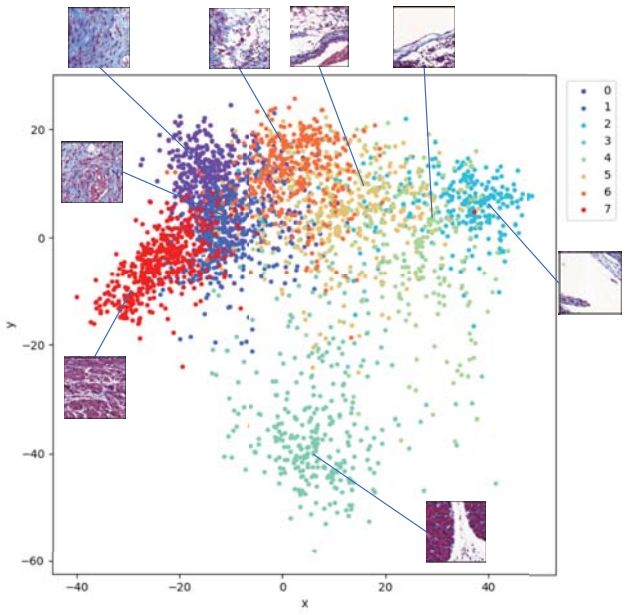

Fig. 2. 2 dimansional space of features

IV. Conclusion

We analyzed histopathological images using IMSAT and obtained interpretable discrete representations. From this results, we are considering the analysis of mice created under other experimental conditions in the future. It has been reported that gene expression subtypes could already be identified with high accuracy in previous studies [10] that utilized supervised learning. However, it becomes difficult to interpret features of clusters when the number is large, moreover, if it is larger thanb 16 or more, gradients vanishing problem occured. Since the misclassification has increased as the cluster size increases, It seems that it was classified by the brightness of the whole image rather than the detailed structure of the tissue. Therefore, We need to perform more detailed clustering even with the same color distribution among samples. We concluded that it is necessary to modify the model so that it can be distinguished even when the lesion site is small.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 17H05297.

References

[1] Daisuke Komura and Shumpei Ishikawa, “Machine Learning Methods for Histopathological Image Analysis,”Computational and Structural Biotechnology Journal, vol. 16, pp.34–42, 2018.
[2] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii., “Virtual adversarial training: a regularization method for supervised and semi-supervised learning,”arXiv preprint arXiv:1704.03976, 2017.
[3] Krause, Andreas and Pietro Perona and Ryan G. Gomes, “Discriminative Clustering by Regularized Information Maximization,”Advances in Neural Information Processing Systems, vol. 23, pp.775–783, 2010.
[4] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama, “Learning discrete representations via information maximizing self augmented training,”arXiv preprint arXiv:1702.08720, 2017.
[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, “Intriguing properties of neural networks,”arXiv preprint arXiv:1312.6199, 2013.
[6] Cover, Thomas M and Thomas, Joy A. Elements of information theory. John Wiley & Sons, 2012.
[7] John S. Bridle and Anthony J. R. Heading and David J. C. MacKay, “Unsupervised Classifiers, Mutual Information and 'Phantom Targets,”Advances in Neural Information Processing Systems, vol. 4, pp.1096–1101, 1992.
[8] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian, “Deep residual learning for image recognition,”CVPR, pp.770–778, 2016.
[9] J. B. Tenenbaum, V. de Silva, J. C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,”Science, vol. 290, pp.2319–2323, 2000.
[10] Coudray, Nicolas and Moreira, Andre L and Sakellaropoulos, Theodore and Fenyo, David and Razavian, Narges and Tsirigos, Aristotelis, “Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images using Deep Learning,”Nature Medicine, vol. 24, pp.1559–1567, 2018.