Metastatic cancer detection from histopathology images using convolutional neural network

Wei Wei, Omer Kose-Shahul, Wangzhuo Shi and Qiaosi Tang
Georgia Institute of Technology

Abstract

Histopathology examination of tissue samples is an essential tool in the detection and diagnosis of metastatic cancer. In this project, we trained several CNN architectures, such as DenseNet, Resnet, Xception, Inception, VGG, compare their individual performances and the performance of the ensemble classifier to detect metastatic cancer cells, with abnormal cell morphology. We employed data augmentation, such as image rotation, flipping, to increase available training data and robustness of the classifier. We visualized CNN feature maps and gradients to understand the decision making behind the classification and studied randomized position of location of the metastatic cancer pixel to understand if positional bias exists in our model. Altogether, this study will contribute to the digital pathology field in assisting histopathology detection with high accuracy and consistency.

1. Introduction/Background/Motivation

The spread of cancer, also known as metastasis, is a key characteristic of cancer and is the primary contributor to cancer-related fatalities worldwide. Detecting cancer metastasis and delivering effective and timely treatment is key to improved treatment outcomes. To determine whether cancer has metastasized beyond the initial tumor site, assessing the biopsy of sentinel lymph node, which is the first lymph node that cancer cells are likely to reach, is a common diagnostic procedure. In this procedure, the removed sentinel lymph nodes are sectioned into thin slices, stained, and imaged by microscope for histopathology analysis. Conventionally, the histopathology analysis is performed by a pathologist, whereby abnormal cellular structures such as cancer cells are manually detected and reported. Although this approach is widely practiced, the manual assessment is time-consuming, subjective, and the results can be variable among different pathologists. Given the pivotal role of histopathological assessment in cancer detection and the limitations in manual practice, developing computational tools for automated cancer detection from histopathological images is in strong demand.

In recent years, deep learning approaches such as convolutional neural networks (CNNs) have become a robust tool for image-based cancer diagnosis, ranging from radiology imaging to histopathology tissue analysis. These efforts provide a more accurate and efficient solution to assist healthcare professionals, thus contributing to early cancer detection and improved patient outcomes. With this motivation, in this study, we aim to develop a prototype of CNN-based cancer detection tool for sentinel lymph node histopathology image classification. We evaluated the performance of six CNN models as well as several data preprocessing strategies. The models are trained and assessed on a subset of the PatchCamelyon (PCam) benchmark dataset (Figure 1) [1,2]. This dataset consists of 220,025 histopathology image patches (equal size of 96x96 pixels, RGB channels) from lymph node biopsies with manual binary annotations of whether each image contains cancerous tissue (labeled with 1) or not (labeled with 0). With this dataset, we simplified the metastatic cancer cell detection task to a binary classification task. Because in clinical practice, high-resolution histopathological images are typically divided into small image patches and each image patch will be labeled with or without cancerous tissue, our developed approach using this patch-level dataset can be transferable to identify cancerous regions within a larger histopathology image for real-world applications.

2. Approach

In this study, we applied the ResNet-50/52, EfficientNet-B0/B7, GoogLeNet, VGGNet16, DenseNet201, Xception pre-trained CNN architectures (Figure 2) for binary classification of the histopathology images. All these model architectures were pre-trained on ImageNet with 1.2 million natural images of 1,000 classes. We fine-tuned these models on the PCam dataset for histopathological classification task. To improve the robustness of the classifier, we combined several data augmentation methods and evaluated their effects. We also studied if the pretrained weights in all layers can improve the validation accuracy or if all these layers need to be retrained with updated weights for higher validation accuracy. We anticipated explosion of trainable parameters with challenging compute and training time. We encountered challenges in hyper parameter tuning with or without data augmentation parameters. We tried to manage these issues with selective application of Batch normalization and Flatten layers to reduce the complexity without compromising accuracy of models. We also tried trainable vs non-trainable layers in pre-trained models to reduce the number of trainable parameters. We tried to take advantage of the feature extraction capabilities of the pre-trained layers of these various models for training that last layer, so that reduced number of parameters are needed to train the model.

2.1. ResNet-50/52

Residual Network (ResNet) was proposed by He et al., and has achieved the state-of the-art performance on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [3]. The pre-trained ResNet has been used for transfer learning in various domains. For example, ResNet-50 has been applied to classify benign or malignant tumor in BreakHis, a breast cancer histopathology image dataset [1]

2.2. EfficientNet-B0

EfficientNet was first developed by Tan and Le in 2019 and has demonstrated state-of-the-art accuracy on image classification tasks while being computationally efficient [2]. It presents a novel compound scaling approach that simultaneously scales the depth, width, and resolution of the neural network [2]. In the domain of histopathology image classification, EfficientNet architectures (B0-B7) have been used to classify breast cancer histopathology images [4]. In this study, we fine-tuned ResNet-50 (Figure 2A) and EfficientNet-B0 (Figure 2B) for 10 epochs on 80% of the PCam dataset and validated on the remaining 20%. PyTorch framework was used to implement the pre-trained models. The training and validation were performed on NVIDIA Tesla V100 GPU.

2.3. GoogLeNet

GoogLeNet is a variant of the Inception Network developed by a team at Google in 2014 (aka Inception V1). The inception module uses 1x1 convolution to effectively reduce model dimensions. It has 9 inception modules fit linearly, in total 22 layers deep or 27 layers if including the pooling layers. The training was done in total 20 epochs, 10 with freezing and 10 with unfreezing of the pre-trained CNN layer weights.

2.4. VGGNET16

VGGNET16 or OxfordNet is a CNN-based model introduced by Simonyan and Zisserman from Oxford university in an article titled "Very deep convolutional networks for large-scale image recognition" [3]. VGGNET16 uses a series of 3x3 kernel filters and nonlinear layers which significantly improves feature learning and classification performance and is an ideal candidate for learning complicated patterns involved in histopathology tissue analysis. Our evaluation included a pretrained VGGNET16 as a model for feature extraction. One 512-dimensional fully connected layer is added after the final max-pooling layer and one dropout layer to prevent overfitting and reduce training time. Figure 7 illustrates the architecture of the VGGNET16 network. TensorFlow and Keras framework are used for Data Augmentation using ImageDataGenerator. RandomSearch is used for hyperparameter tuning. VGGNET16 achieved 0.91 accuracy rate in validation runs.

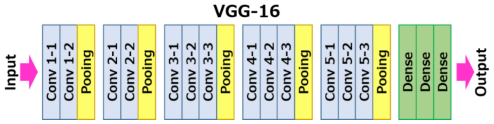

Figure 1. VGGNET16 Architecture

We tested with pre-trained weights of VGGNET16 with trainable flag set to False and found the validation accuracy around 0.81. When we trained all weights in all VGGNET16 layers, the accuracy jumped to 0.91. We hypothesize because cancer images involve complex patterns, retraining pre-trained models makes model extract features and improve detection accuracy. Data augmentation with added rotation, shift, flip and zoom improved accuracy. Figure 3 illustrates AUC, Precision and Recall evaluation metrics. Our conclusion is VGGNET16 is one of the top contenders for cancer detection with high quality images with simple tweaking of dense and dropout layers.

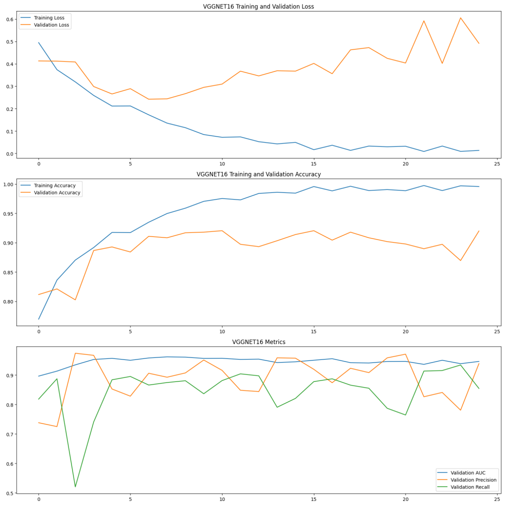

Figure 2. VGGNET16 Training/Validation loss/accuracy

Hyper parameters		Data Augmentation (ImageDataGenerator)		Metrics (Validation)	
Hyper parameters		Data Augmentation (ImageDataGenerator)		Metrics (Validation)	
epochs	25	rotation_range	40	loss	0.5159
batch_size	64	width_shift_range	0.2	accuracy	0.8891
train_steps	8000	height_shift_range	0.2	AUC	0.9333
val_steps	2000	shear_range	0.2	precision	0.89
Learning_rate	0.0001	zoom_range	0.2	Recall	0.8221
Optimizer	Adam	horizontal_flip	TRUE		
Loss function	binary cross entropy	fill_mode	nearest		
Drop out	0.3				
Validation accuracy	0.914499998				

Table 1. VGGNET16 Hyper Parameters/Data Augmentation/Metrics .

2.5. DenseNet201

Densenet201 contains direct connections from any layer to all subsequent layers. nth layer receives the feature maps of all preceding n-1 layers. Figure 4 illustrates the architecture of DenseNet201.

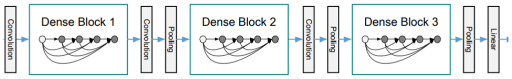

Figure 3. DenseNet201

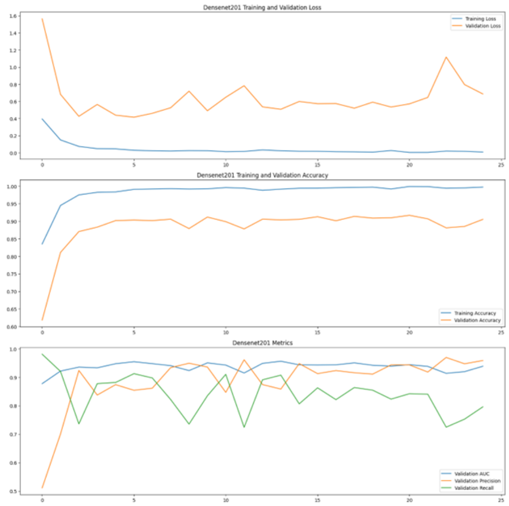

Figure 4. DenseNet201

We used hyper parameters as in Figure 5 and compared both frozen layers of DenseNet201 against fully trained model with data augmentation. The performance of fully trained DenseNet201, on both accuracy and precision, outperformed with accuracy 0.9052 the one with previous layers being frozen (transfer learning). While fully trained DenseNet201 achieved high accuracy of classification, we found the cost of training is higher. The denser the blocks being retrained, the greater accuracy, and it is because previous layers can learn specific features and these features are lost when previous layers are frozen. One 512-dimensional fully connected layer is added after the final max-pooling layer and one dropout layer to prevent overfitting and reduce training time.

2.6. Xception

The Xception model [10] is a deep convolutional neural network (CNN) architecture that was introduced by François Chollet, the creator of the Keras deep learning library. Xception stands for ”Extreme Inception,” and it is inspired by the Inception architecture. The key innovation in Xception is the use of depth wise separable convolutions, which are more computationally efficient than traditional convolutions. This allows the model to achieve better performance with fewer parameters. In our study, we train the full dataset with Xception as base model in a similar setup(train-validation split, hardware, etc.) to EfficientNet-B0.

2.7. Data augmentation

Given the smaller size of PCam dataset compared to ImageNet used for pre-training, we anticipated the problem of overfitting. To prevent overfitting, in addition to choosing network architectures with low complexity (such as using EfficientNet with B0 scale), we applied a serial transformation technique to the input data. We first applied a random choice of color jittering, followed by a random choice of image rotation, horizontal or vertical flip, or a combination of rotation and flip. We performed an ablation study to demonstrate that our data augmentation strategy can improve the robustness of classifier.

Models	Data augmentation	AUROC
EfficientNet-B0	Yes	0.9075
EfficientNet-B0	No	0.8648
ResNet-50	Yes	0.8977
ResNet-50	No	0.836
Xception	Yes	0.8497
Xception	No	0.8396

Table 2. Testing set AUROC with/without data augmentation.

3. Experiments and Results

3.1. Experimental Setups

In this study, we applied transfer learning to the pre-trained ResNet-50, EfficientNet-B0, VGGNET16 and DenseNet201 architectures. The last layers of the fully connected layer of the pre-trained architectures are adjusted on output dimension for binary classification. The pre-trained ResNet-50, EfficientNet-B0, VGGNET16 and DenseNet201 models were independently trained with 10-25 epochs, with or without data augmentation.

3.2. Optimizer

For ResNet-50, EfficientNet-B0, VGGNET16 and DenseNet201 trainings, we used Adam as the optimizer. Parameters excluding the final fully connected layer were applied with a learning rate of 0.0005, and the final fully connected layer was applied with 10 times higher learning rate of 0.005. The optimizer is configured with a weight decay of 0.0001.

3.3. Loss function

For ResNet-50, EfficientNet-B0, VGGNET16 and DenseNet201 trainings, Cross-Entropy Loss (CELoss) was used. To address the overfitting and overconfidence problems, label smoothing was performed as a regularization technique with the smoothing factor at 0.1. For Googlenet, binary cross entropy loss (BCEloss) was used.

3.4. Evaluation metrics

We evaluated ResNet-50, EfficientNet-B0, VGGNET16 and DenseNet201 performances by accuracy and loss on the allocated training and validation sets. In addition, predictions performed on a separate test dataset from Kaggle were evaluated by AUROC. For Googlenet, F-metric was used.

3.5. Data augmentation alleviated overfitting

To assess the effects of data augmentation in prediction performance, we performed an ablation study on ResNet-50, EfficientNet-B0 and Xception models to train with or without the data augmentation strategy (section 2.3). In the training and validation datasets, when data augmentation is not applied, although both models demonstrate higher accuracy and lower loss in training and validation, we observed a large gap between training and validation. This indicates that without data augmentation, the models are impacted by overfitting and cannot generalize well. The overfitting effect is alleviated after data augmentation is applied. Furthermore, as demonstrated in Table 2, in the testing dataset, applying data augmentation can improve prediction performance in all three models mentioned above, which again confirmed that data augmentation can mitigate overfitting.

3.6. Gradient visualization verified decision making

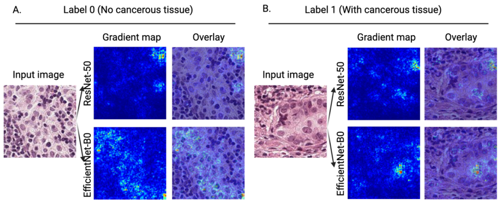

Figure 5. Gradient map visualization of ResNet-50 and EfficientNet-B0

To understand the decision making behind the classification, we visualized the gradient maps of ResNet-50 and EfficientNet-B0 trained with data augmentation. As shown in Figure 4, we compared the gradient maps of input images with cancerous tissue versus those without. Although ResNet-50 and EfficientNet-B0 attend to the input image differentially, the most high-magnitude gradients are focused on regions with cellular component, indicating the validity of decision making. Particularly, EfficientNet-B0 was able to focus on a region with abnormally enlarged nuclei, which has been considered as one of the morphological hallmarks of cancer cells [7]. We can verify model decision-making through gradient visualization.

3.7. The Effect of Training Centered Cancer Pixel Images on Test Image Predictions (Googlenet)

We noticed that the cancer pixels in this dataset are described as localized in the center 32x32 patch of the images. We want to explore whether models trained on images with centered cancer pixels show bias when predicting on images whose cancer pixels are off-center.

For this experiment, we used Googlenet with the following transformations: Center crop to reduce input images from 96x96 to 64x64 and then resized to the image size required for Googlenet: 224x224. Predictions were made on 2 sets of 1000 identically sampled images, with the 1st set transformed in the same transformations as the model training samples and the 2nd set transformed using random crop instead of centered crop, and then rotated 90 degrees.

We found that, indeed, the Googlenet model trained on centered cancer pixel made different predictions on some of the same images: comparing the predicted labels from the 1st set and the 2nd set, we found that the ”accuracy” was 95%. In other words, 5% of the same images had different labels that resulted from how the images were transformed. This calls into question what features the deep neural networks are learning to classify cancer pixels, and whether deep neural networks are memorizing the input images instead of learning image features.

As an extension to this experiment, we tried to replicate the findings using Xception with the following transformations and further locate the locational bias in the Pcam dataset: Center crop to reduce the input images from 96x96 to 32x32 and then rescale the images(np.repeat) back up to 96*96 to be accepted by Xception model.

Predictions made on the competition test set show inconsistency with our given information that cancer pixels are localized in the center 32x32 area. Reasons that could lead to this include: 1) Human Errors. 2) Xception model fails to extract features from these center pixels to make classification decisions.

3.8. Model Interpretation (Googlenet) and Visualization of input image features

To visualize pixels in an input image that contributed to the classification as cancer pixel, we used Saliency and Integrated Gradients methods from Captum to visualize an input image that was labeled as cancer (see Figure 5). Inception5b is the last CNN layer module of Googlenet. We used Layer Gradcam method from Captum to visualize the gradient of the target with respect to Inception5b convolution layer (see Figure 6).

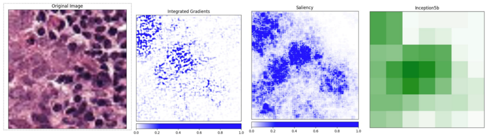

Figure 6. Saliency and Integrated Gradient Maps of Cancer Image and Layer Gradcam Attribution of Inception5b CNN layer with Googlenet

A histopathologist will be more familiar with the interpretation of the resulting images using Saliency and Integrated Gradient methods for input image and Layer Gradcam for last CNN layer, we see pixels and positions in the Inception5b CNN layer corresponding to higher values that are differentially stained in the original image, in addition to the enlarged nuclei as described with Efficientnet and Resnet visualizations (10 points) How did you measure success? What experiments were used? What were the results, both quantitative and qualitative? Did you succeed? Did you fail? Why? Justify your reasons with arguments supported by evidence and data.

3.9. Competition Test Set Scoring using Ensembled Labels

For this group project, the team members have each individually trained 1-2 deep neural network models and generated labels for the competition test set, which contains 57,458 images. One of common tricks to improve test set score is to use majority vote and see if ensemble model has improved performance than each individual models. We found in our specific case, the performance of the ensemble model is not better than the best of the individual models, 0.89 vs 0.9075 with Efficientnet (see Table 3).

4. Conclusion

Experimenting with multiple pre-trained neural network architectures, Efficientnet, Resnet, Googlenet, Vggnet, Densenet, and Xception, we were able to achieve scores on the competition test images in the range of 82% to 91%. In one reported comparison between deep neural networks scoring of histopathological melanoma images vs those of 11 pathologists, CNN achieved a mean sensitivity/specificity/accuracy of 76%/60%/68%, outperforming the mean sensitivity/specificity/accuracy of 51.8%/66.5%/59.2% from the humans [13]. Even though we do not have the performance metrics on this test set from expert histopathologists, we would like to think the ¿90% performance scores on the 57,458 images using the CNNs are respectable. We have visualized input images and layers to attribute labeling of cancer pixels. With Efficientnet and Resnet, cancer pixels correspond to regions of enlarged nuclei, with Googlenet, cancer pixels correspond to differentially stained areas.

5. Work Division

Please see the table 4 4 for the team member contributions.

References

[1] Q. A. Al-Haija and A. Adebanjo. Breast cancer diagnosis in histopathological images using resnet-50 convolutional neural network. 2020. 2020 IEEE International IOT, Electron-

Model Name	Competition Test Set Performance
	(Public Score)
Efficientnet (w/ transforms)	0.9075
Resnet50 (w/ transforms)	0.836
Googlenet (w/ center cropping transforms)	0.8597
Vggnet16	0.8549
Densenet201	0.8219
Xception	0.8497
Ensemble model using majority vote	0.89

Table 3. Individual Model Performance on Competition Test Set vs Ensemble Model.

Student Name	Contributed Aspects	Details
Qiaosi Tang	Efficientnet, Resnet	Data augmentation, visualization .
Wei Wei	Googlenet	Effect of centered cancer pixel, visualization, ensemble scoring .
Omer Kose-Shahul	Vggnet, Densenet	Data Augmentation, visualization, Latex Report .
Wangzhuo Shi	Xception	Data augmentation, effect of centered cancer pixel .

Table 4. Contributions of team members.

ics and Mechatronics Conference (IEMTRONICS), Vancouver, BC, Canada, 2020, pp. 1-7, doi: 10.1109/IEMTRONICS51293.2020.9216455. 2

[2] Q.V. Le. M. Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. 2020. Proceedings of ICML. 2

[3] Zisserman A Simonyan K. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint, http://arXiv.org/abs/1409.1556. 2

[4] Y.Nagaraju Venkatesh, R.K.Sheela and D.A.Sahu. Histopathological image classification of breast cancer using efficientnet. 2022. 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-8, doi: 10.1109/INCET54531.2022.9824351. 2