Detecting Cancer Metastases on Gigapixel Pathology Images

Final Project of
Applied Deep Learning

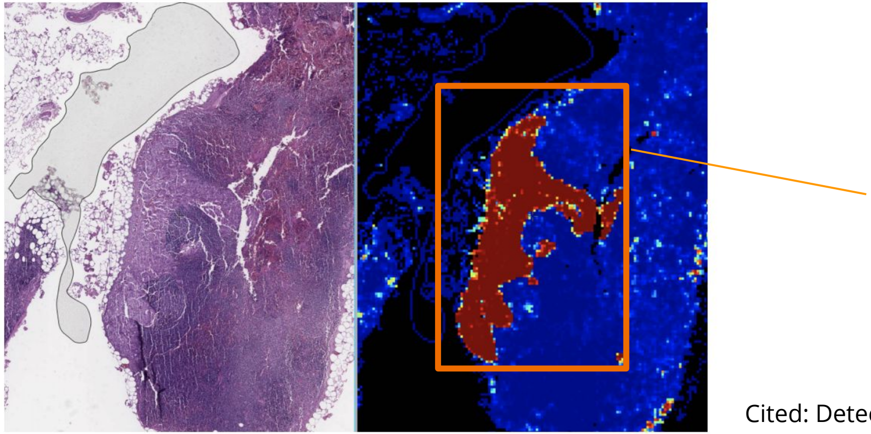
Rui Bai (rb3454) Yichi Liu (yl4327)

Project Introduction

- Motivation
- Flow of the Project

Motivation

- Metastasis detection :
  - Detect whether the breast cancer has spread to nearby cells
  - Early diagnosis will help doctors to give treatment
- However, manually labelling the cell will be time-consuming.
- We designed an automatically Metastasis detection model on the Pathology image with CNN models



Cited: Detecting Cancer Metastases on Gigapixel Pathology Images, 2017

Flow of the Project

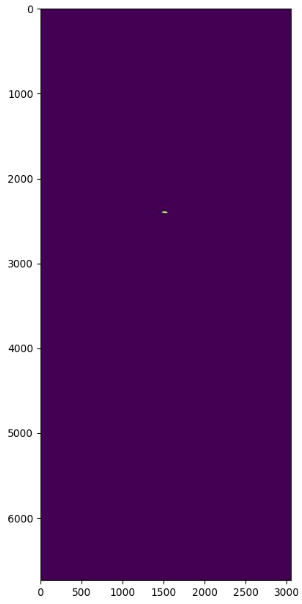
- Project Flow Introduction: introduce the overall steps as below
- Data Processing
  - Training & Validation & Testing
  - Patch extraction for training & validation data
- Model Architectures
  - Different transferred models
  - Different Scales
- Heatmap Construction
- Model Comparison
- Comparison of Results
- Final Prediction for 3 testing data

Data Processing

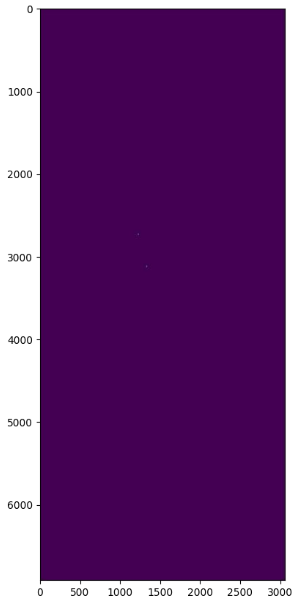
- Training & Validation & Testing
- Patch Extraction for Training & Validation data

Training & Validation & Testing

- Some slides have little cancer cell.
  We don’t take them into consideration
- Training and Validation Image:
  Patches from 8 slides:
  slide 016, 031, 064, 075, 078, 084, 094, 101
  - Training: 80% of patches
  - Validation: 20% of patches
- Testing Image:
  Patches from 3 slides:
  Slide 091, 096, 110
- Observation: **Imbalanced** dataset



Slide 012:
Little cancer cell.
Drop it.

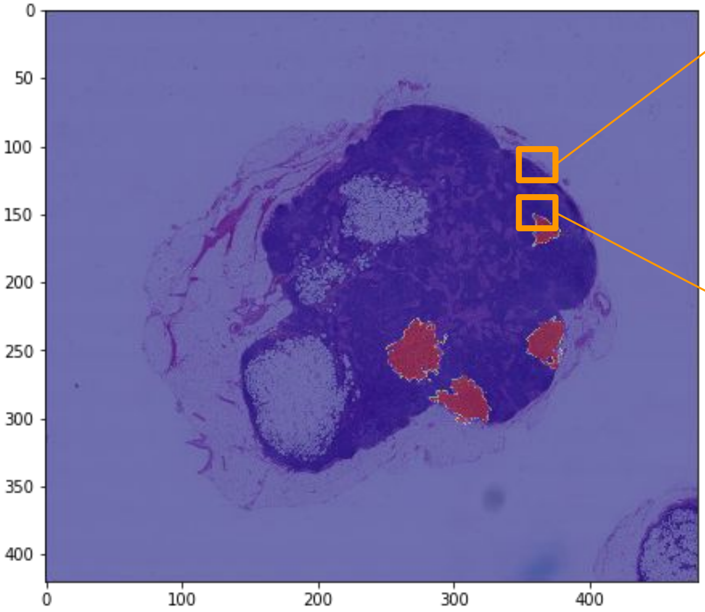


Slide 035:
Almost no cancer cell.
Drop it.

Positive Patch Extraction for Training & Validation data

Randomly Get 200 Positive 299*229 patches:

- If the center point of the patches is not cancer: **Drop it**
- If the center point of the patches is cancer: **Save** the patch and label as **Positive**



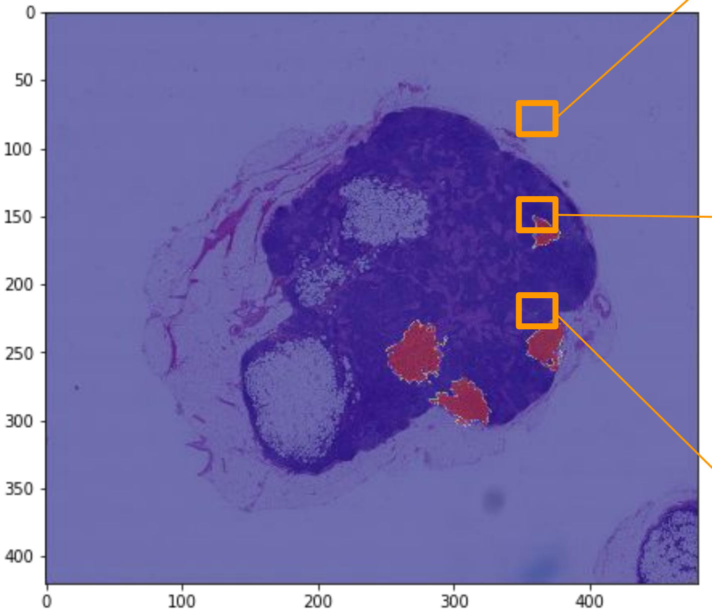
Condition a:
No cancer at the center point.
Solution:
Drop it

Condition b:
The center point is cancer.
Solution:
Label is **POS**

Negative Patch Extraction for Training & Validation data

Randomly Get 200 Negative 299*229 patches:

- If the center point is not tissue (intensity > 0.8): **Drop it**
- If the center point is tissue:
  - If the center region (128*128) contains cancer: **Drop it**
  - If the center region doesn't contain cancer: **Save** the patch and label as **Negative**



Condition a:
No tissues at the center point.
Solution:
Drop it

Condition b:
The center region (128*128) contains cancer.
Solution:
Drop it

Condition c:
The center region (128*128) has no cancer.
Solution:
Label is **NEG**

Avoiding Memory Issues

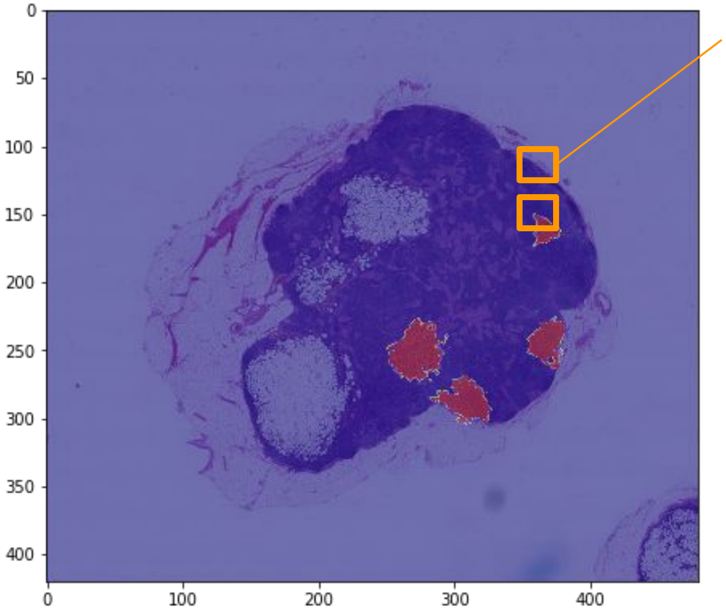
- Don’t Read the whole slide image
- Extract the mask and the slide image only with size 229*229
- Save the patches into folders

We could run our script with even level 0 without crashing in Colab

(without update to Colab Pro)

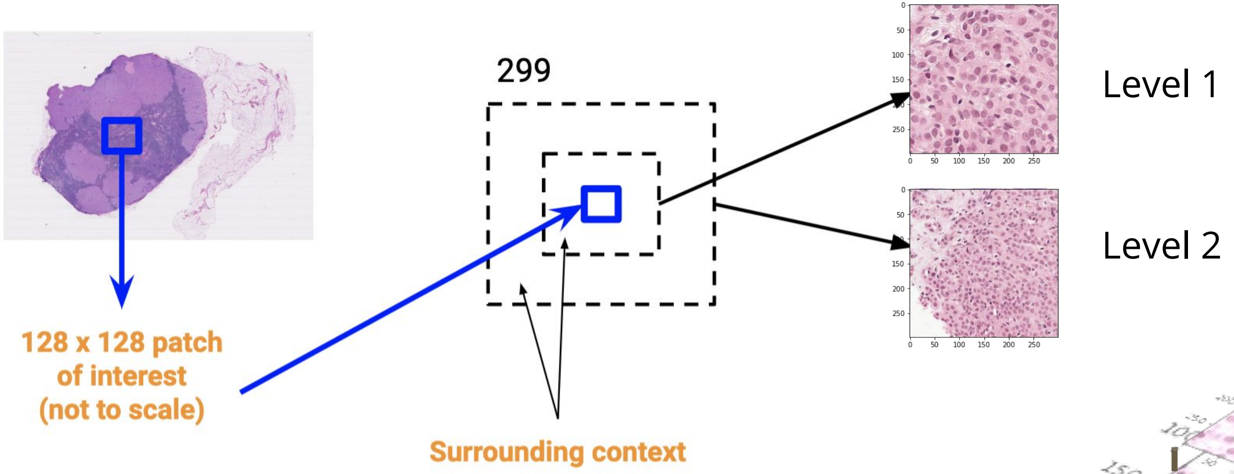
Negative Extraction for Training & Validation data

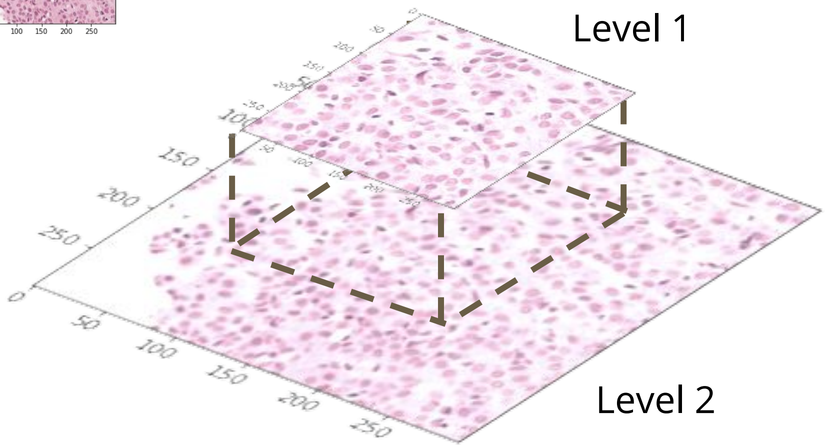
- Randomly Get 200 Positive patches:
  - Conditions:
    - If
- Negative patches:
  - Get tissue pixels
  - Randomly select a tissue pixels to capture a patch
  - Check if tumor in center region
    - If yes: drop it
    - If no: extract this patch



Conditio
n a:

Multi Scale Patch Extraction





For each extracted patch from previous step, get a larger surrounding context by extracting a patch on a higher level, with the same center

Data Augmentation

- Augmentations from the paper:
  - Rotate the patch by 0°, 90°, 180°, 270°
  - Apply a left-right flip and repeat rotations
  - Perturb color
  - Small offset of some pixels

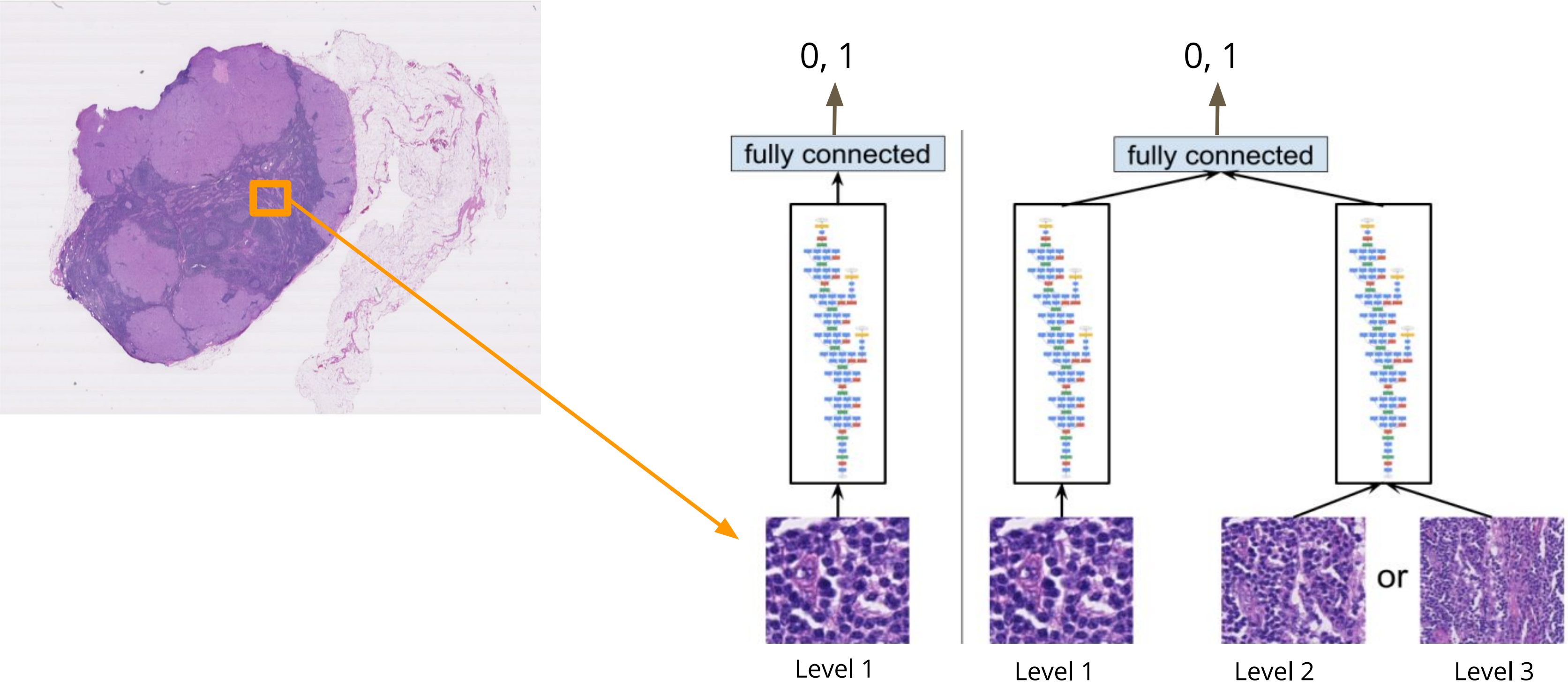
- Additional augmentations we did:
  - Apply an up-down flip, since pathology slides do not have canonical orientations

Model Architectures

Model Architecture

Single-scale

Multi-scale

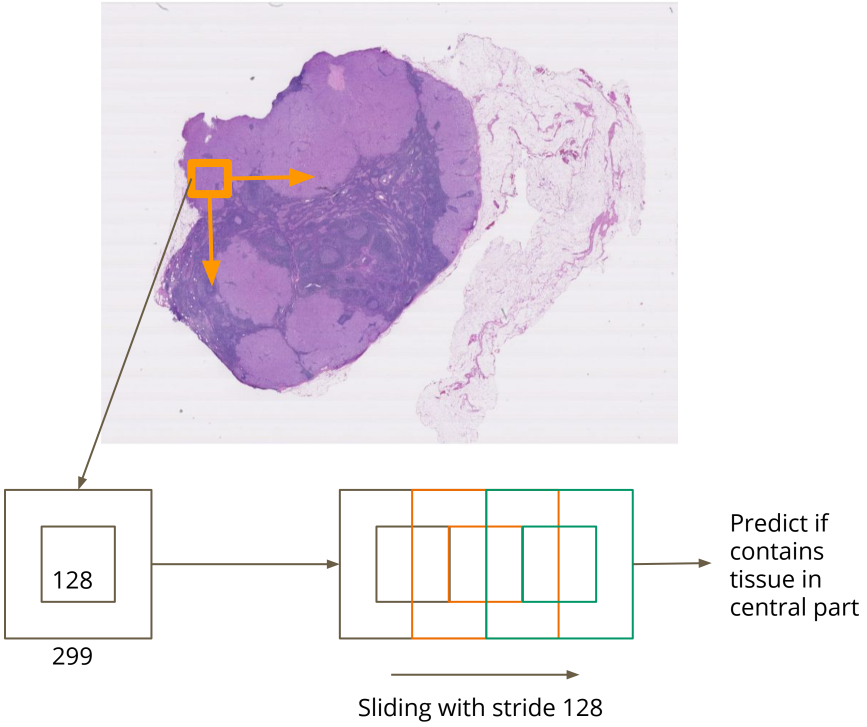


Heatmap Construction

Heatmap Construction Methodology

For each testing slide:

- Sliding a window of size 299*299 through the entire image to extract patches
- Using stride = 128 to match the center region's size, so that the prediction do not overlap
- Predict only If the patch contains tissue in its center 128*128 region



Prediction of Patch

For each patch, we calculate prediction result in two ways:

- Method 1:

  Do a single Prediction on the patch

- Method 2:

  Apply the rotations and left-right flip to obtain predictions for each of the 8 orientations, and average the 8 predictions.

8 Orientations:

Average the 8 predictions

Model Comparison

Single-Scale Models Comparison

Scale Level	Model	Parameters	Validation Accuracy
Level 1	InceptionV3	Transferred	0.9594
		Fine Tuned	**0.9891**
Level 2		Transferred	0.9297
		Fine Tuned	0.9812
Level 3		Transferred	0.9484
		Fine Tuned	0.9703
Level 1	ResNet50	Transferred	**0.80**
		Fine Tuned	**0.82**

Best:
Fine Tuned InceptionV3

Worst:
ResNet50

Thus, we used InceptionV3 model for the following analysis

Multi-Scale Models Comparison

Scale Level	Model	Parameters	Validation Accuracy
Level 1 & 2	InceptionV3	Transferred	0.9869
		Fine Tuned	**0.9906**
Level 2 & 3		Transferred	0.9731
		Fine Tuned	0.9859
Level 3 & 4		Transferred	0.9650
		Fine Tuned	0.9859
Level 1 & 3		Transferred	**0.9516**
		Fine Tuned	0.9641

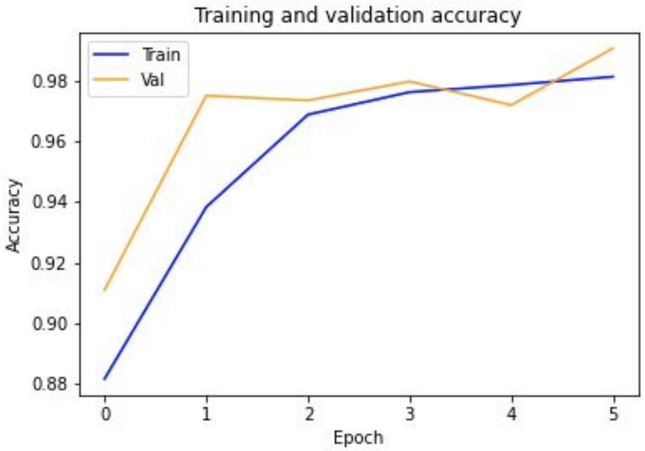
Best:
Fine Tuned InceptionV3
Using Scale Levels 1 & 2

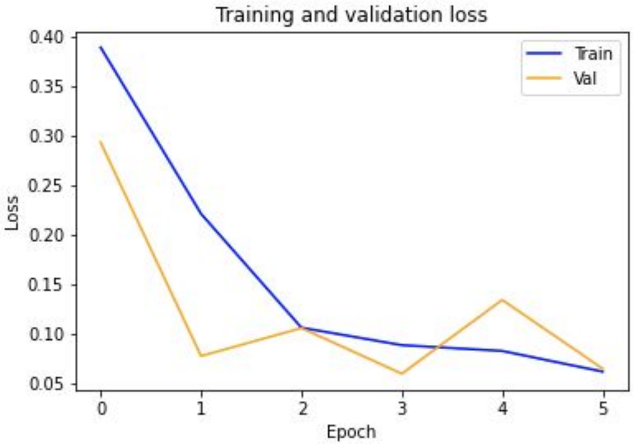
Worst:
Transferred InceptionV3
Using Levels 1 & 3

We used **fined tuned InceptionV3 model** for the following analysis

Model Training Process -- For the best model

- Fine-tuned Interception based model
- Take level 1 and level 2 as input. Level 1 is the reference level that we label the patch.
- We used **early stopping** to prevent overfitting
- **Learning rate** was set to 0.0001

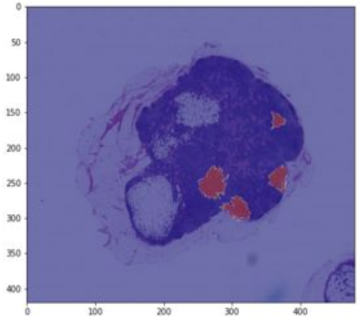




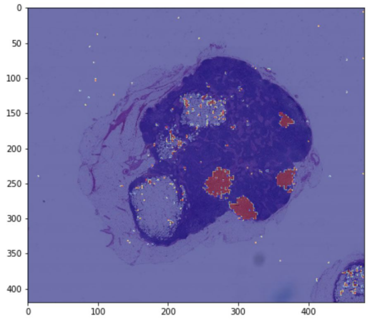
Comparison of Results

This section uses one of the test data, Slide 091, for comparison.

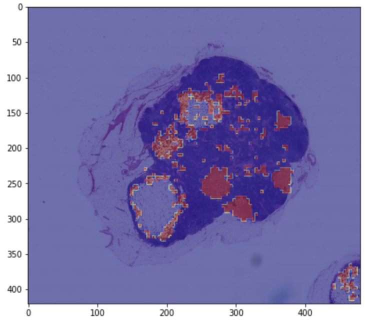
Comparison 1:
Single-scale v.s. Multi-scale



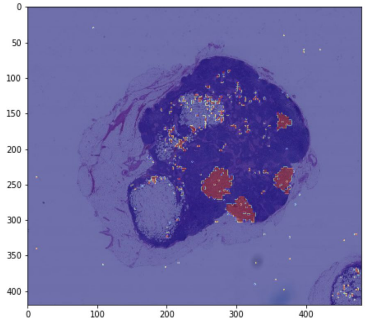
True Label



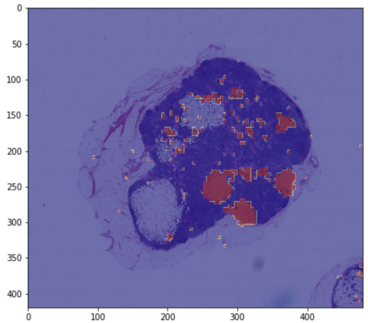
Single-scale Level 1
Precision: 0.65, Recall: 0.72
F1 score: 0.69, AUC score:0.96



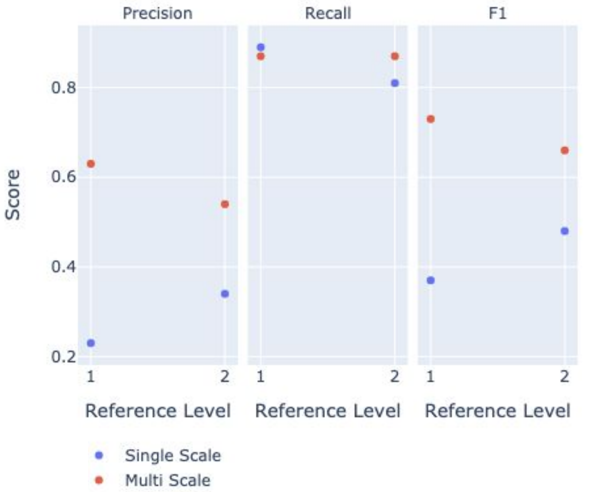
Single-scale Level 2
Precision: 0.34, Recall: 0.81
F1 score: 0.48, AUC score: 0.97



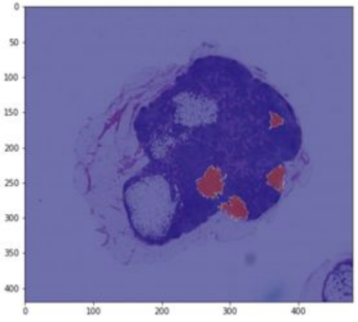
Multi-scale Level 1 & 2
Precision: 0.63, Recall: 0.87
F1 score: 0.73, AUC score: 0.97



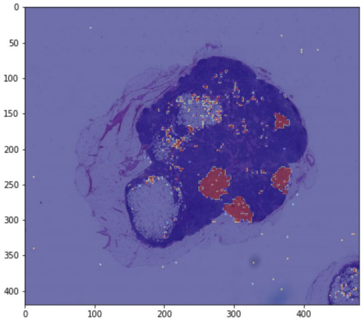
Multi-scale Level 2 & 3
Precision: 0.54, Recall: 0.87
F1 score: 0.66, AUC score: 0.98



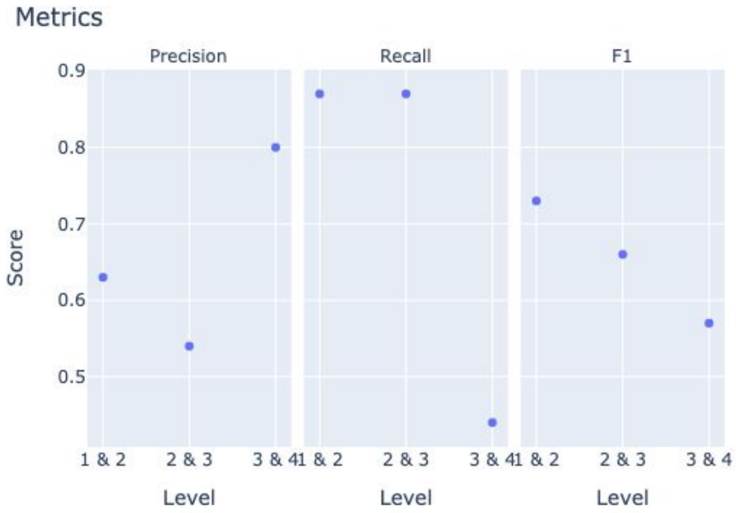
Comparison 2:
Low Levels v.s. High Levels Scales

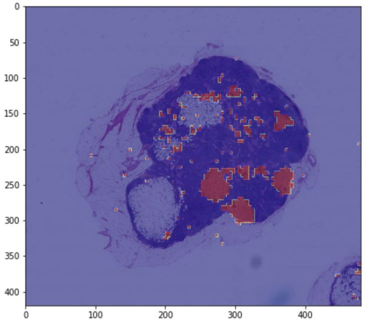


True Label

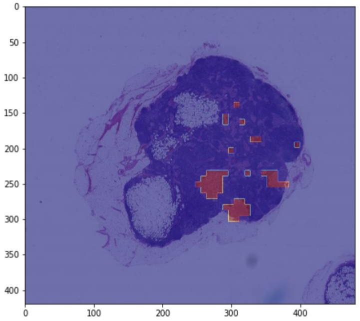


Multi-scale Levels 1 & 2
Precision: 0.63, Recall: 0.87
F1 score: 0.73, AUC score: 0.97



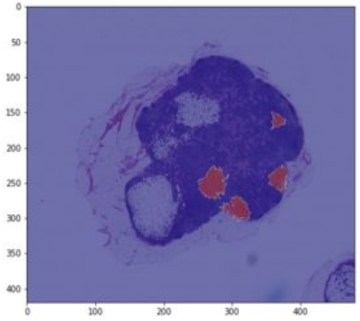


Multi-scale Level 2 & 3
Precision: 0.54, Recall: 0.87
F1 score: 0.66, AUC score: 0.98

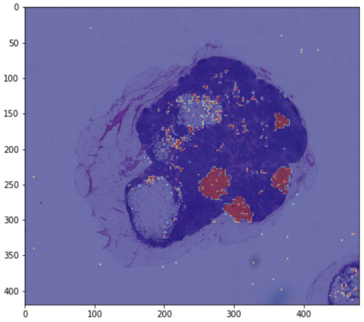


Multi-scale Level 3 & 4
Precision: 0.80, Recall: 0.44
F1 score: 0.57, AUC score: 0.93

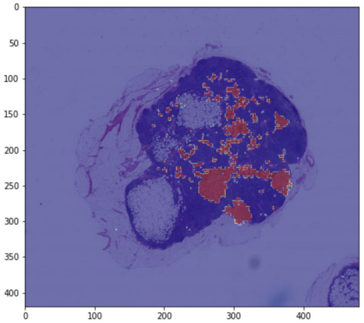
Comparison 3:
Large v.s. Small Surrounding Context



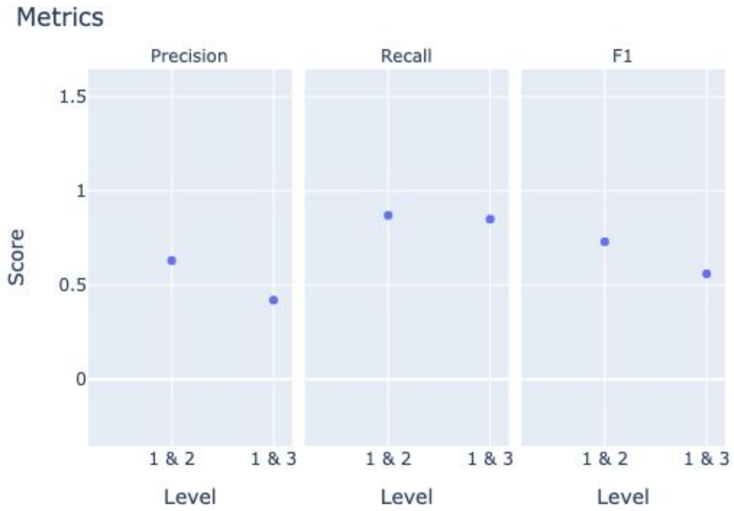
True Label



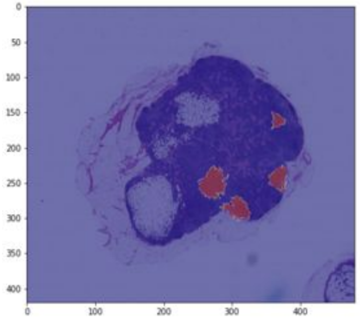
Multi-scale Levels 1 & 2
Precision: 0.63, Recall: 0.87
F1 score: 0.73, AUC score: 0.97



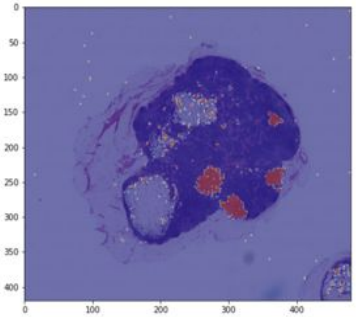
Multi-scale Level 1 & 3
Precision: 0.42, Recall: 0.85
F1 score: 0.56, AUC score: 0.98



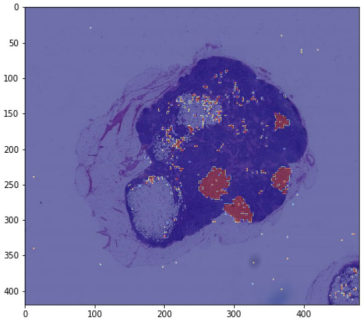
Comparison 4:
Single Prediction v.s. Mean of 8 Predictions Per Patch



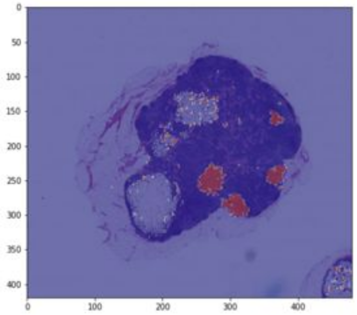
True Label



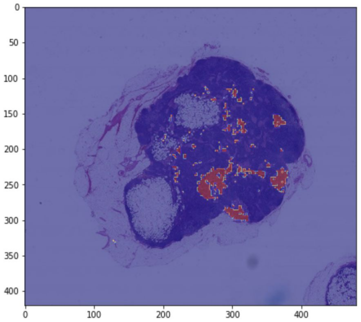
Single-scale Level 1 - Single Prediction
Precision: 0.65, Recall: 0.72
F1 score: 0.69, AUC score: 0.96



Multi-scale Levels 1 & 2 - Single Prediction
Precision: 0.63, Recall: 0.87
F1 score: 0.73, AUC score: 0.97



Single-scale Level 1 - Mean of 8 Predictions
Precision: 0.74, Recall: 0.74
F1 score: 0.74, AUC score: 0.97



Multi-scale Level 1&2 - Mean of 8 Predictions
Precision: 0.60, Recall: 0.58
F1 score: 0.59, AUC score: 0.96

Summary of the Result Analysis

- Multi-scale is better than Single-Scale
- Lower Level is better than Higher Level
- Smaller surrounding area is better than larger surrounding area
- Mean prediction of 8 variations is better than a single prediction of a patch

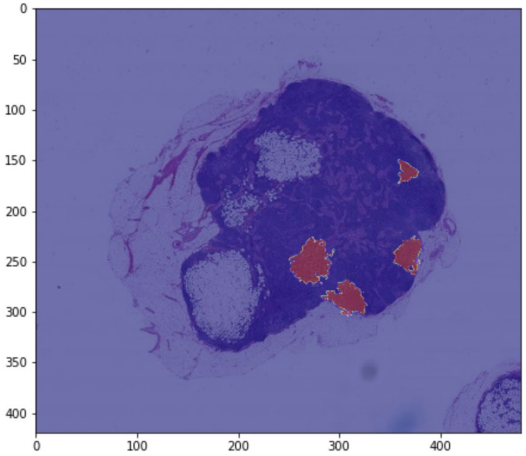
Best Heatmap Predicting Solution:

- Using **multi-scale model of level 1&2**
- Calculating the **mean prediction of 8 variations** of each patch

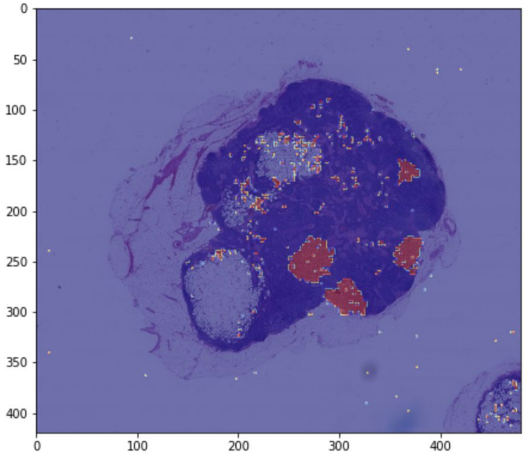
Final Prediction

Predicting the heatmaps for the 3 test slides

Final Predicted Heatmaps on Slide 091

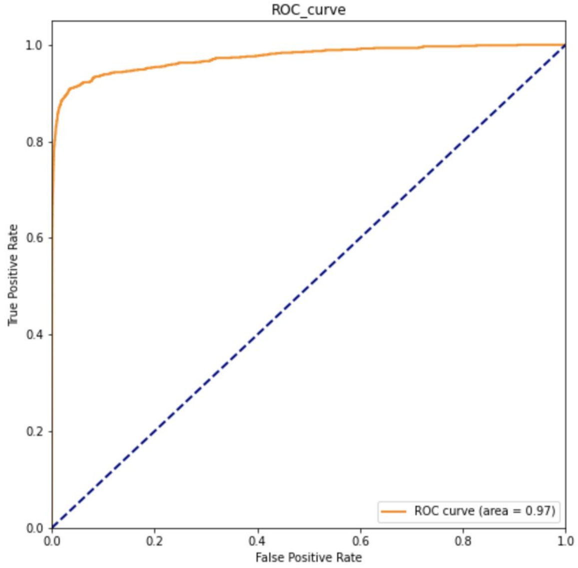


True label



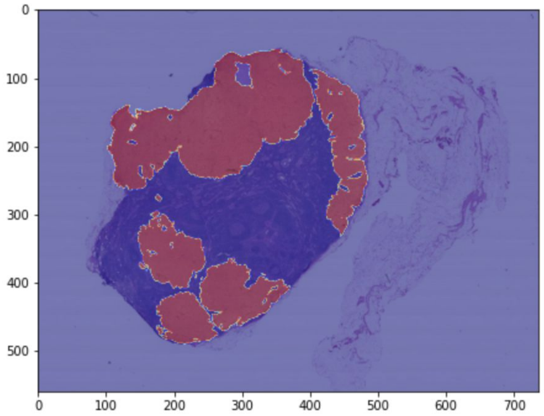
Predicted using Multi-scale model with level 1&2

Precision score: 0.63
Recall score: 0.87
F1 score: 0.73
AUC score: 0.97

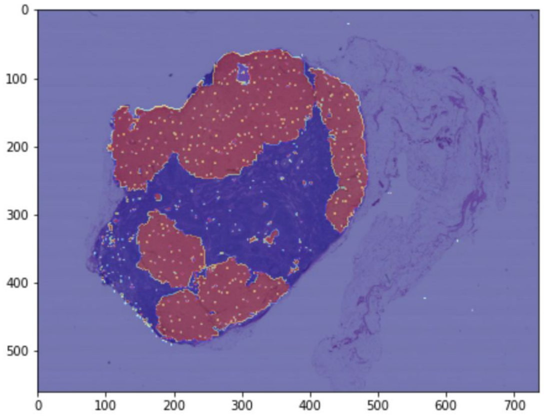


Final Predicted Heatmaps on Slide 110

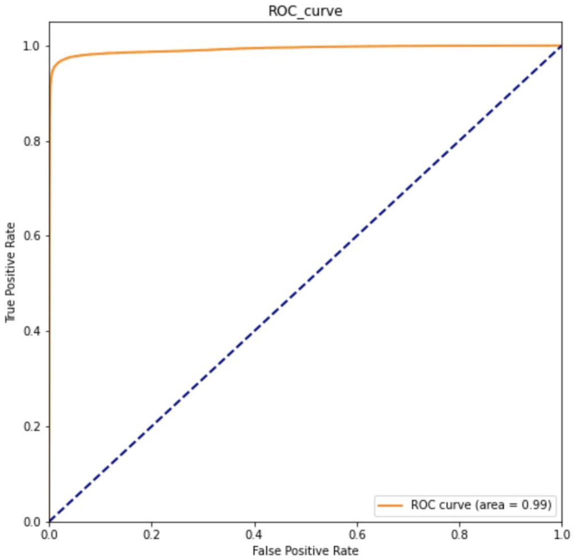
Precision score: 0.96
Recall score: 0.96
F1 score: 0.96
AUC score: 0.99



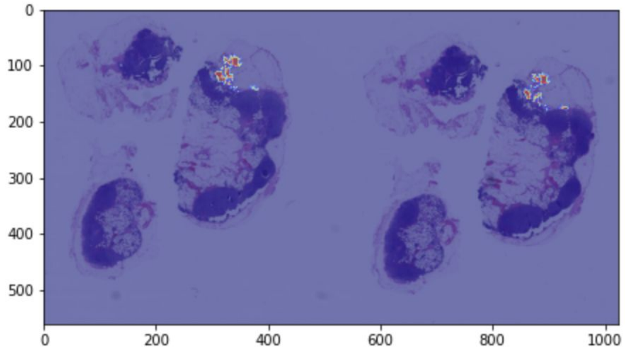
True label



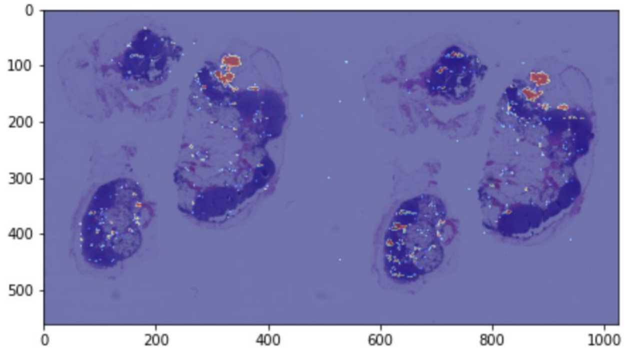
Predicted using Multi-scale model with level 1&2



Final Predicted Heatmaps on Slide 096

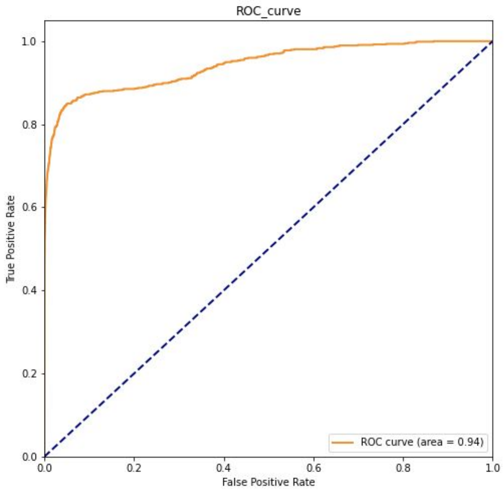


True label



Predicted using Multi-scale model with level 1&2

Precision score: 0.38
Recall score: 0.73
F1 score: 0.50
AUC score: 0.94



Thanks for Watching!

Logic

- Project flow introduction: introduce the overall steps as below
- Data Processing
  - Training & Validation & Testing
    - Some slides has only a little tumor, we dropped them
    - 3 for testing 8 for training and validation
  - Patch extraction for training & validation data
- Model Architectures
  - Transferred
    - InceptionV3 unfix (fine tuned)
    - InceptionV3 unfix (fixed)
    - ResNet unfix (fine tuned)
  - Scales
    - Single Scale: level 1,2,3
    - Multi-scale: level 1&2, 2&3, 3&4, 1&3
- Heatmap Construction
  - Sliding window to extract patches that contain tissue in center part
  - Each patch predict 1 time v.s. predict 8 times?
- Model comparison (表格)
  - Val acc (excel 前两个sheet)
    - Single-scale
    - Multi-scale
  - 结论：用inceptionV3，从这之后不show ResNet
- Model comparison w.r.t. heatmaps of 091 (next page)
- Final heatmaps for 3 testing data (091, 110, 096) using the best model

Model comparison w.r.t. heatmaps of 091

- Single or Multi (3 列点点，横轴是ref level)
  - Inception level 1 v.s. Inception level 1&2
  - Inception level 2 v.s. Inception level 2&3
  - Inception level 3 v.s. Inception level 3&4 不要了
  - 结论：multi的比single好
- Low or high ref level
  - Inception level 1&2 v.s. 2&3 v.s. 3&4
  - 结论：low ref level更好
- How large surrounding area to consider
  - Inception level 1&2 v.s. 1&3
  - 结论：adjacent的更好
- Predict 1 time or 8 times?
  - Inception level 1: 1 time v.s. 8 times
  - Inception level 1&2: 1 time v.s. 8 times
  - 结论：八次的更好
- Final model:
  - Inception level 1&2 unfix 8 times