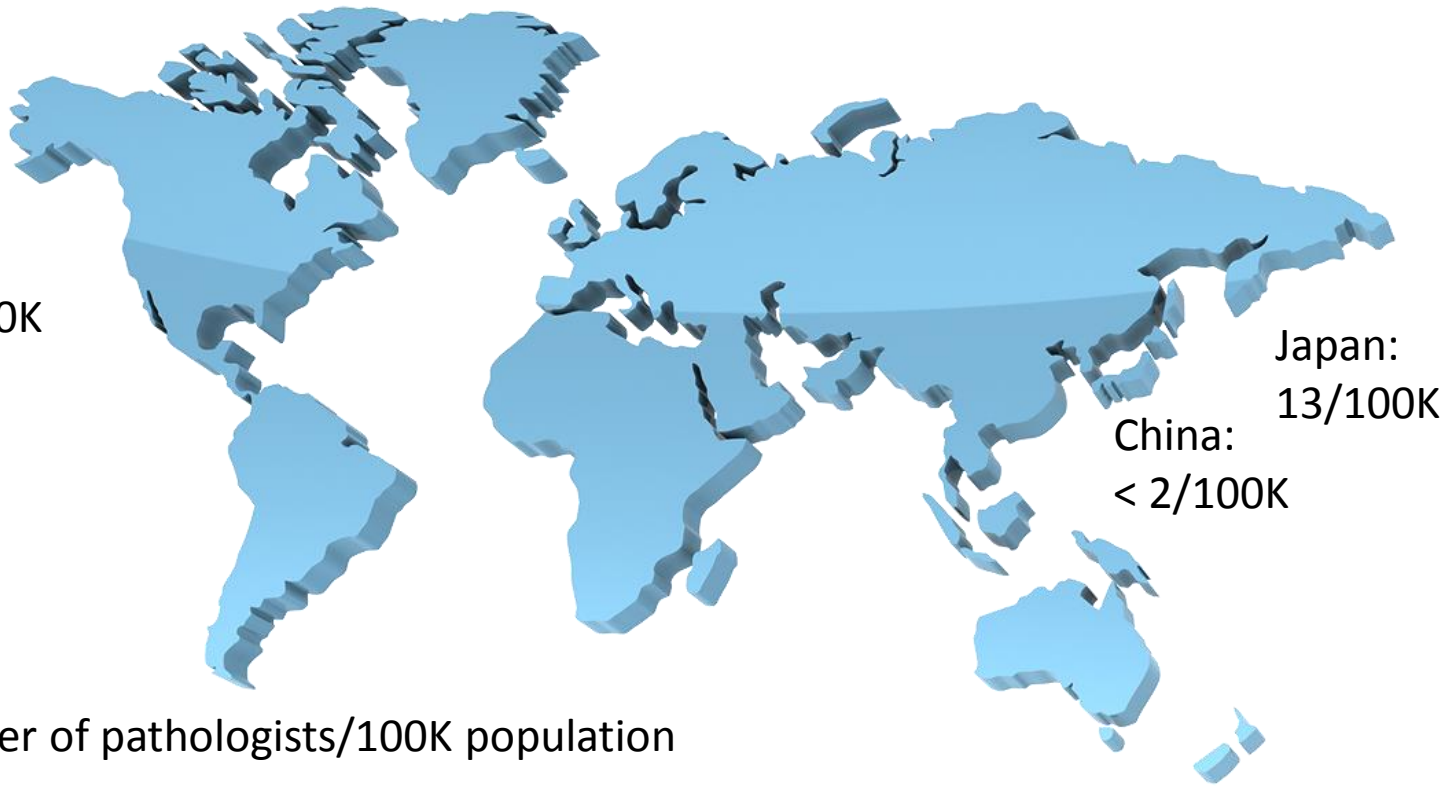
Large-scale Histopathology Image Analysis for Colon Cancer on Azure

Yan Xu[1, 2]
Tao Mo[2]
Teng Gao[2]
Maode Lai[4]
Zhuowen Tu[2,3]
Eric I-Chao Chang[2]

[1]Beihang University; [2]Microsoft Research Asia; [3]UCSD; [4]Zhejiang University

Motivation

- Histopathology assessment the gold standard for cancer detection
- Shortage of trained pathologists



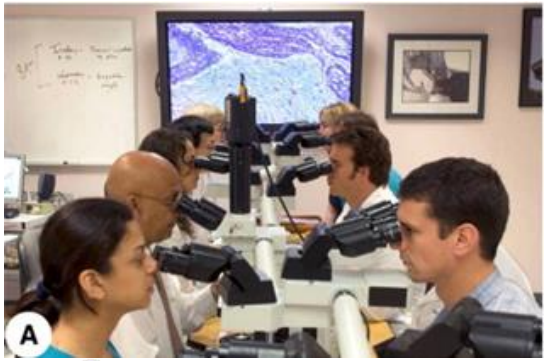
Number of pathologists/100K population

http://www.moh.gov.cn/publicfiles/business/htmlfiles/zwgkzt/ptjnj/year2009/t-2.htm
http://www1.gifu-u.ac.jp/~patho1/250.html

生物与医学工程学院
BUAA SCHOOL OF BIOLOGICAL SCIENCE AND MEDICAL ENGINEERING

Microsoft Research
微软亚洲研究院

Motivation



An expert

10 minutes

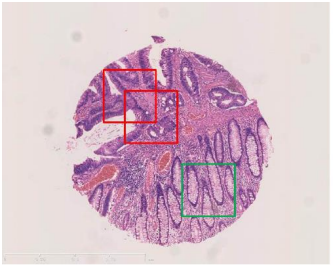
10-20 slides

Introduction

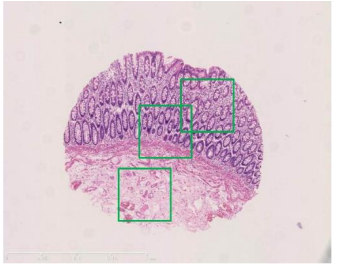
Clinical tasks for histopathology image analysis		
diagnosing the presence of cancer (**classification**)	segmenting cancer cells (**segmentation**)	**clustering** the tissue cells into various sub-classes

Approaches

Unsupervised image segmentation methods may not work well due to their complicated patterns	Most existing **supervised** approaches for tissue cell segmentation require detailed manual annotations	**Weakly supervised** methods use coarse-grained labeling to aid automatic exploration of fine grained information



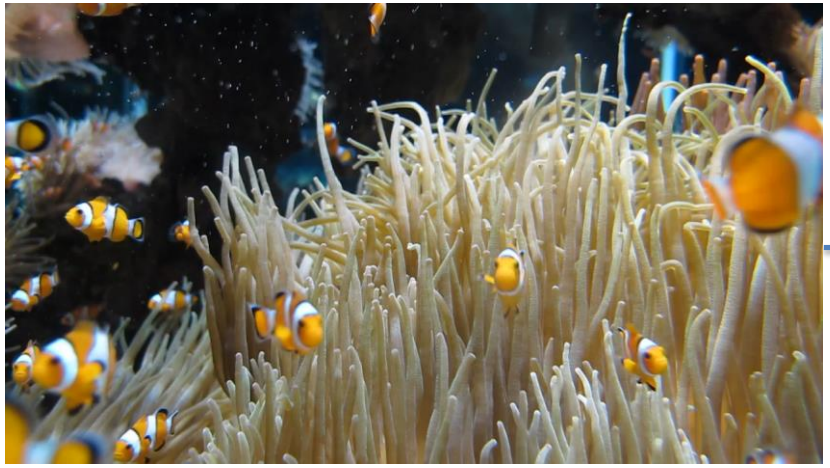
(a) cancer image



(b) non-cancer image

Multiple Instance Learning

Positive Instances



Negative Instances

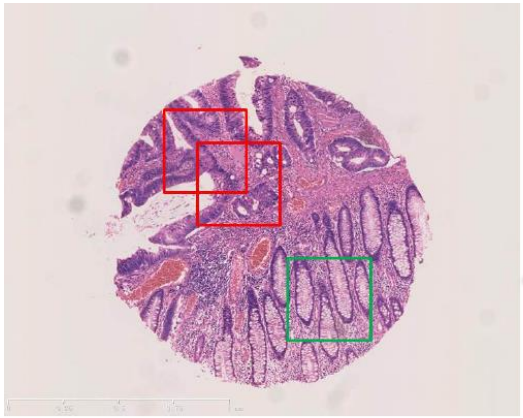
MIL Training

Cardinal Fish Classifier

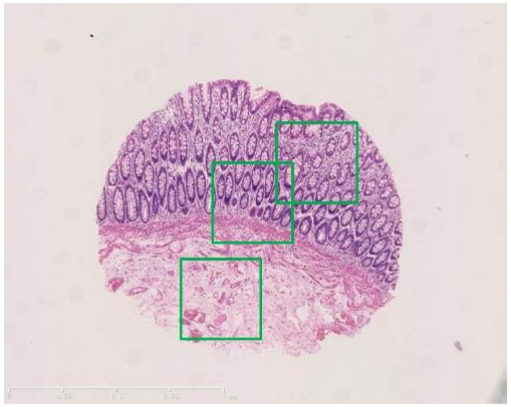
Weakly labeled training data

(Maron 1997, Viola 2005)

Multiple Instance Learning



(a) cancer image



(b) non-cancer image

Positive Instances

Negative Instances

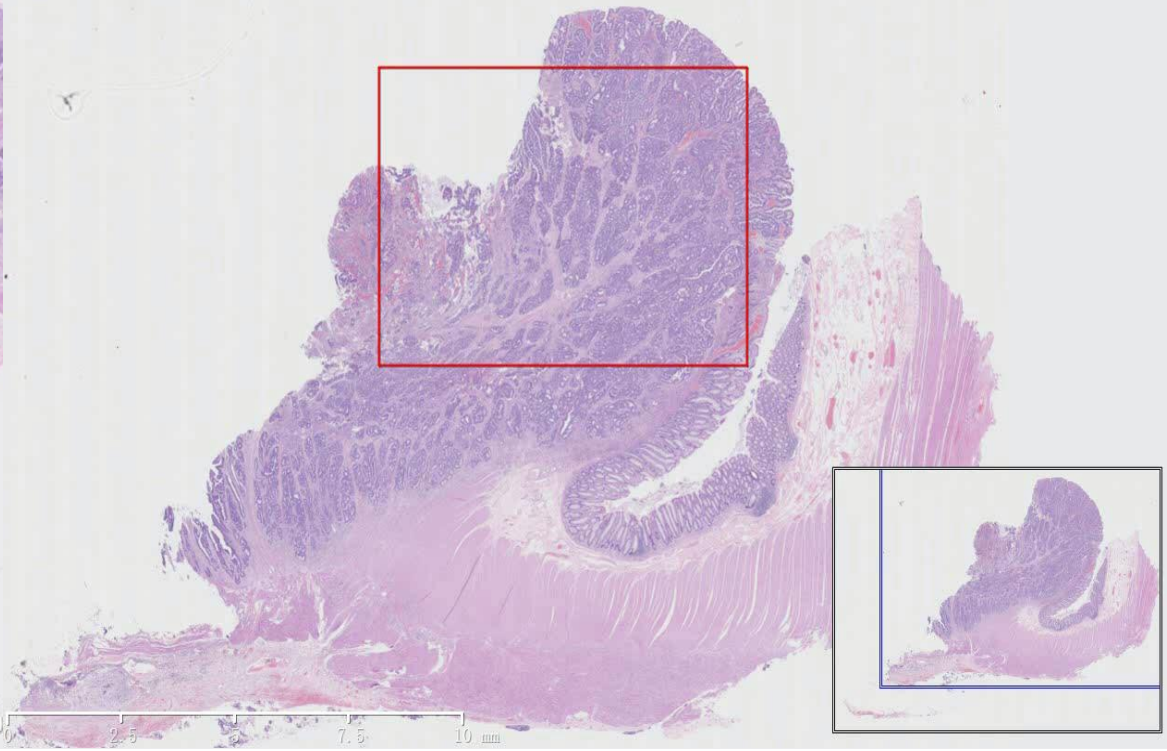
MIL Training

Cancer Image Classifier

Weakly labeled training data

Parallel Multiple Instance Learning

- A standard histopathology slice
  Resolution: 200,000 x 200,000
- Most existing medical imaging tools infeasible
- A single machine with limited memory and disk space



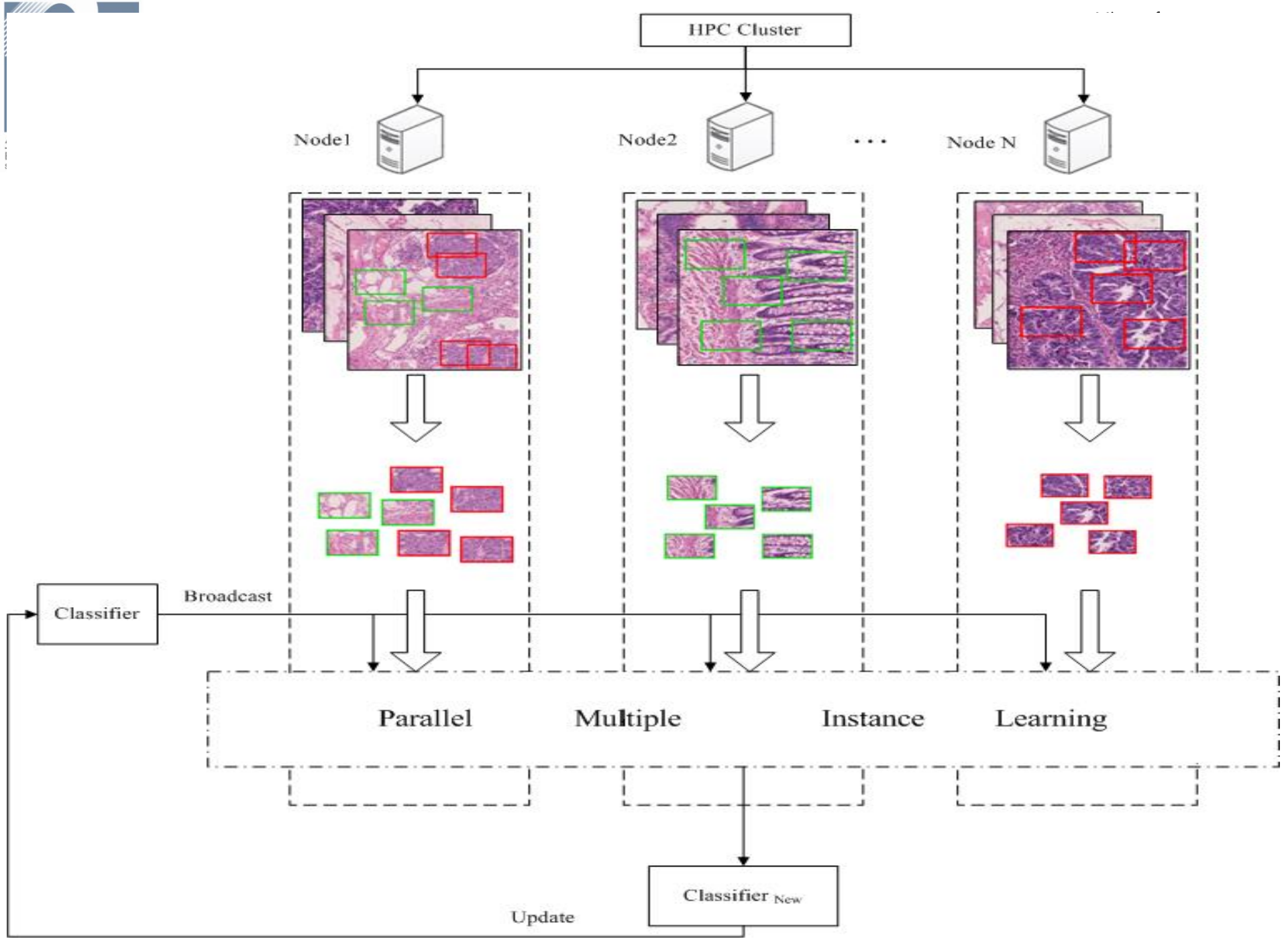
Introduction

- The origin of parallel multiple instance learning (PMIL)

MIL: previous MIL-based work performed classification but not segmentation (Liang 2007, Liu 2010, Dundar 2010)

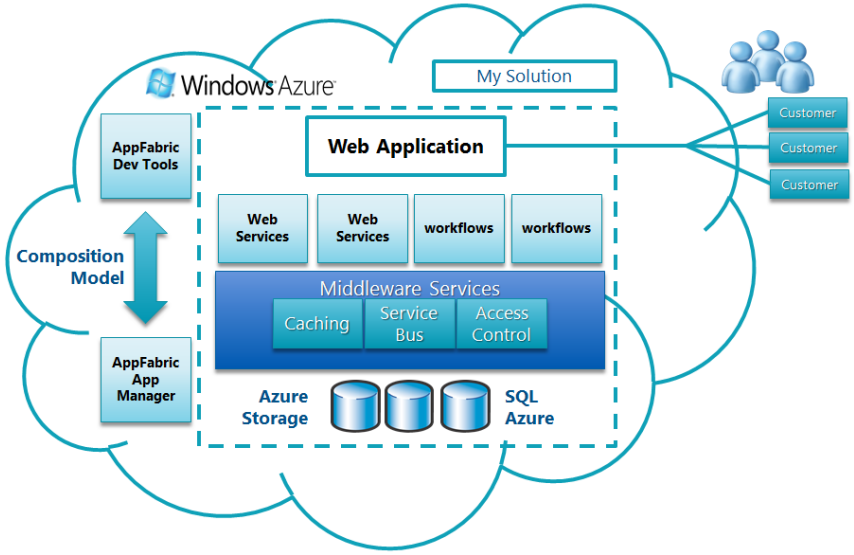
MCIL: adopts the clustering concept into MIL but takes the assumption of independent instances (Xu 2012)

PMIL: emphasizes the parallel concepts and max-margin concept among different clusters, demonstrating the efficiency and effectiveness


HPC Cluster
Node1
Node2
Node N
Classifier
Broadcast
Parallel
Multiple
Instance
Learning
Classifier New
Update

Experiment Setting

- Microsoft Windows HPC Cluster
- 128 compute nodes
  - 2 quad-core Xeon 2.43-GHz processors
  - 16GB RAM
  - 1Gbit Ethernet adapters
  - 1.7TB local disk storage



生物与医学工程学院
BUAA SCHOOL OF BIOLOGICAL SCIENCE AND MEDICAL ENGINEERING

Microsoft Research
微软亚洲研究院

Experimental Setting

Datasets

- 1,118 images at a resolution of 10 billion pixels, subsampled 22,484 pieces (10,000 X 10,000), **6.135 trillion byte colon cancer images**
- Obtained from the Department of Pathology of Zhejiang University using Hamamatsu Nano Zoomer 2.0HT digital slice scanner (40 X)
- Each image is labeled by two pathologists, the third pathologist moderates their discussion

MTA—Moderately or well differentiated tubular adenocarcinoma
PTA—Poorly differentiated tubular adenocarcinoma
MA—Mucinous adenocarcinoma
SRC—Signet-ring carcinoma

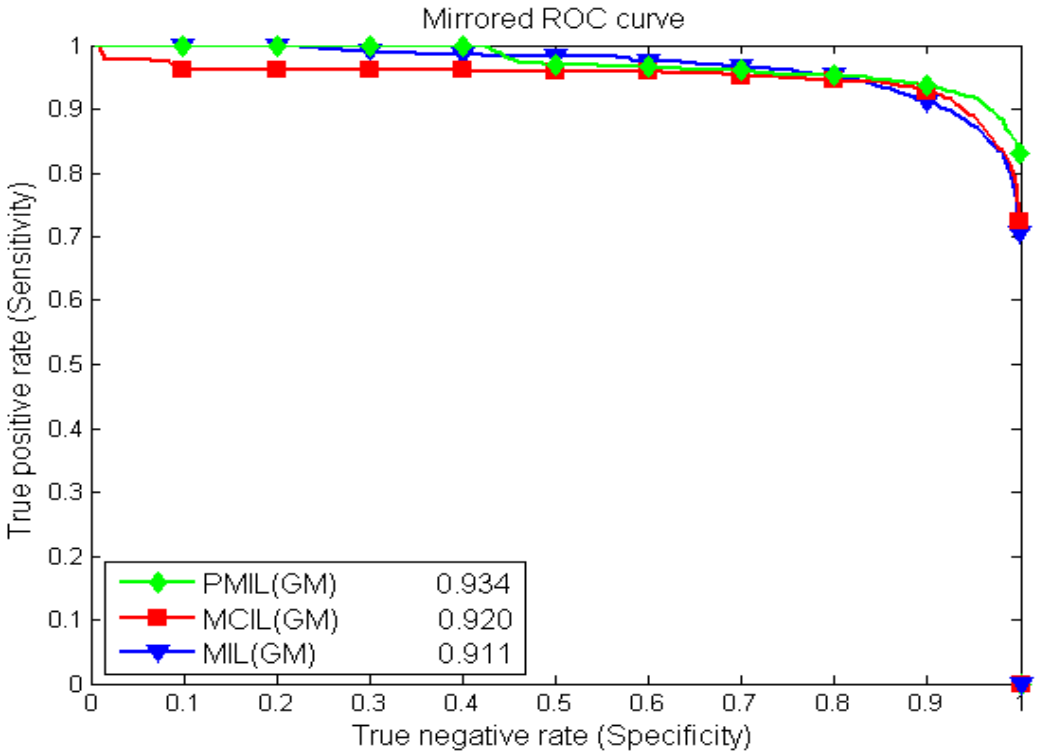
Experimental Setting

Experiment settings

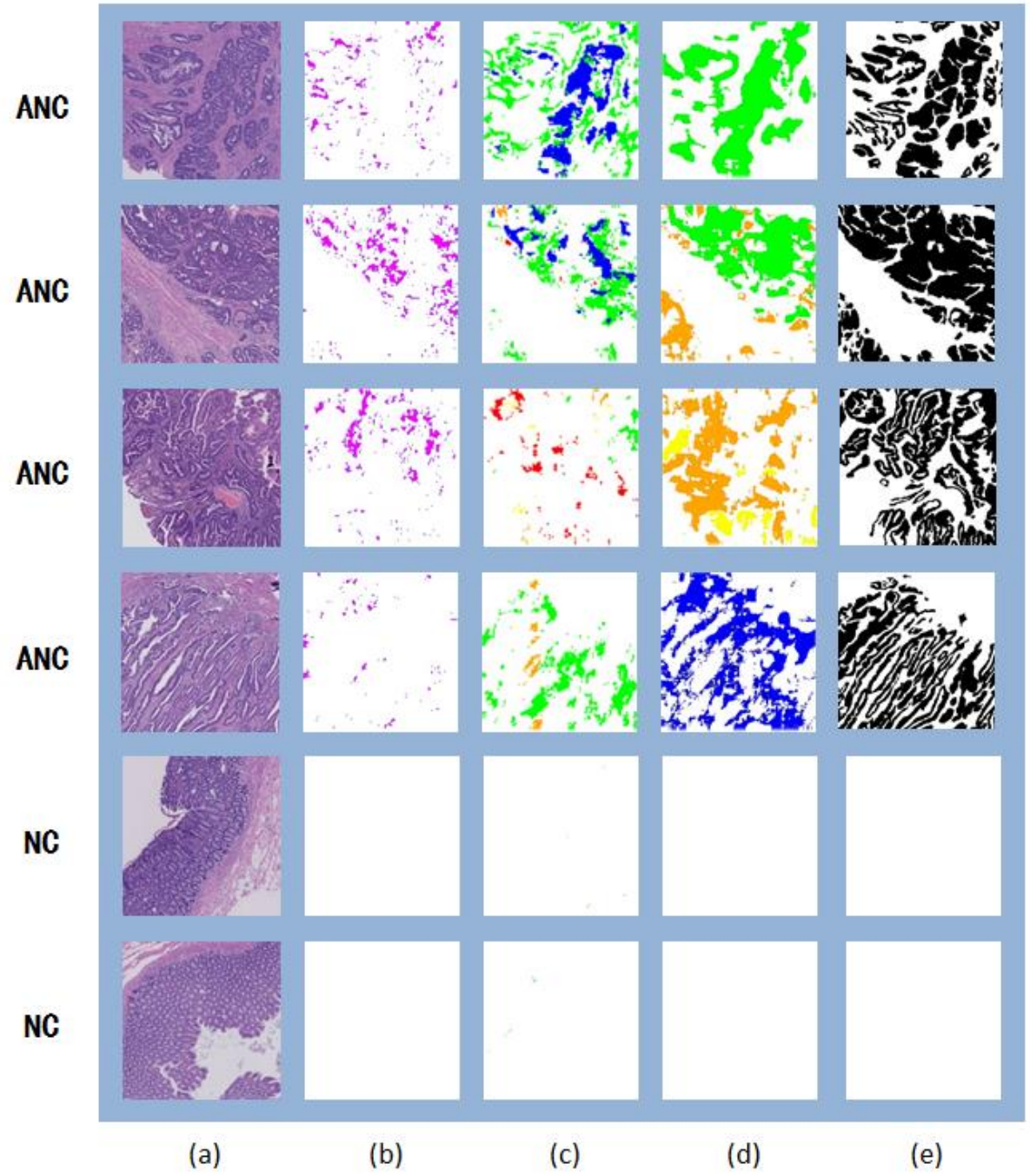
Preprocessing	Down-sample patch extraction (patch size: 640*640), overlap step 100
Parameters	r=20, K=5, T=1000 set the prior weight for each bag as the same value
Features	Common computer vision features are used: L*a*b* Color Histogram, LBP, SIFT, HOG, and GLCM
Others	6-fold cross validation Use Gaussian function as weak classifier

Results

- Results:



Three methods are compared in this experiment with the same features and parameters
Trained using million instances and 215 features in 128 compute nodes (1024 CPU cores) for 25.1 hours



(a) Original images (b) MIL (c) MCIL (d) PMIL (e) Ground truth

Results

- Results:

Compared with other ***weakly supervised methods***

- 64 nodes to generate patch feature vector (D = 215) of 200 million extracted patches (30 hours)
- 128 nodes to train MIL, MCIL and P-MIL models on the training data (T = 1000), it takes 3.2, 24.8 and 25.1 hours respectively

PMIL significantly improves results by competition between clusters.

Results

- Results: Scalability

50 iterations

Cores	NC[s]	C [s]	Overall[s]	Speedup
128	330	32781	33111	1.000
256	408	16413	16821	1.968
512	482	8025	8507	3.892
1024	515	4128	4643	7.131

Non-computation

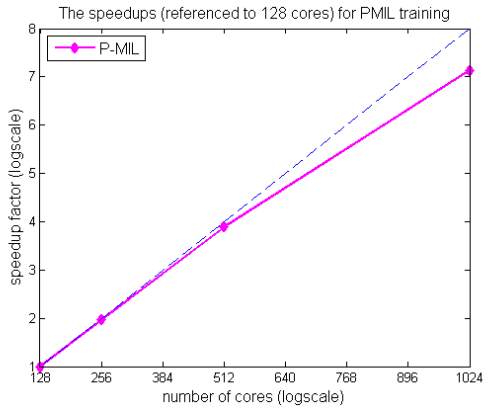
Non-computation (NC) makes up of communication and synchronization between compute nodes

Computation

Computation (C) time is the sum of time except non-computation time

Data distribution

The data distribution stage is excluded because this data I/O heavily depends on bandwidth of disk and network



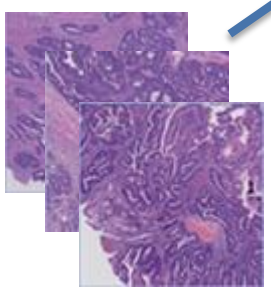
生物与医学工程学院
BUAA SCHOOL OF BIOLOGICAL SCIENCE AND MEDICAL ENGINEERING

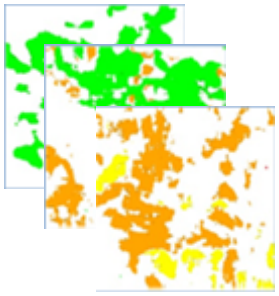
Microsoft Research
微软亚洲研究院

On Microsoft Azure

Parallel Multiple Instance Learning Classifier

On Microsoft Azure





Result:

1. **Cancer or not;**
2. **Cancer location;**
3. **Cancer sub-type;**

Conclusion

Parallel Multiple Instance Learning

By introducing the parallel multiple instance learning framework using MPI/multi-threading hybrid programming model on Windows HPC cluster

We choose HPC as parallel computing platform is that **Microsoft Azure** added support for HPC and MPI, which enables our algorithm to scale up to cloud with minor porting effort

Microsoft Research
微软亚洲研究院

Thanks!

xuyan@buaa.edu.cn