Breast Cancer Screening with Deep Neural Networks

Krzysztof J. Geras

Joint work with **Kyunghyun Cho**, **Linda Moy**, **Gene Kim**, **Stacey Wolfson** and **Artie Shen**

Neural networks

They are amazing.

Neural networks

They are amazing. Can we save the world?

Breast cancer screening

Breast cancer screening

- Data are X-ray images. Basically, an image recognition problem.

Breast cancer screening

- Data are X-ray images. Basically, an image recognition problem.
- Problem with a high impact on the society.

Breast cancer screening

- Data are X-ray images. Basically, an image recognition problem.
- Problem with a high impact on the society.
  - In 2015, 232k women in the US were diagnosed with breast cancer, 40k died.

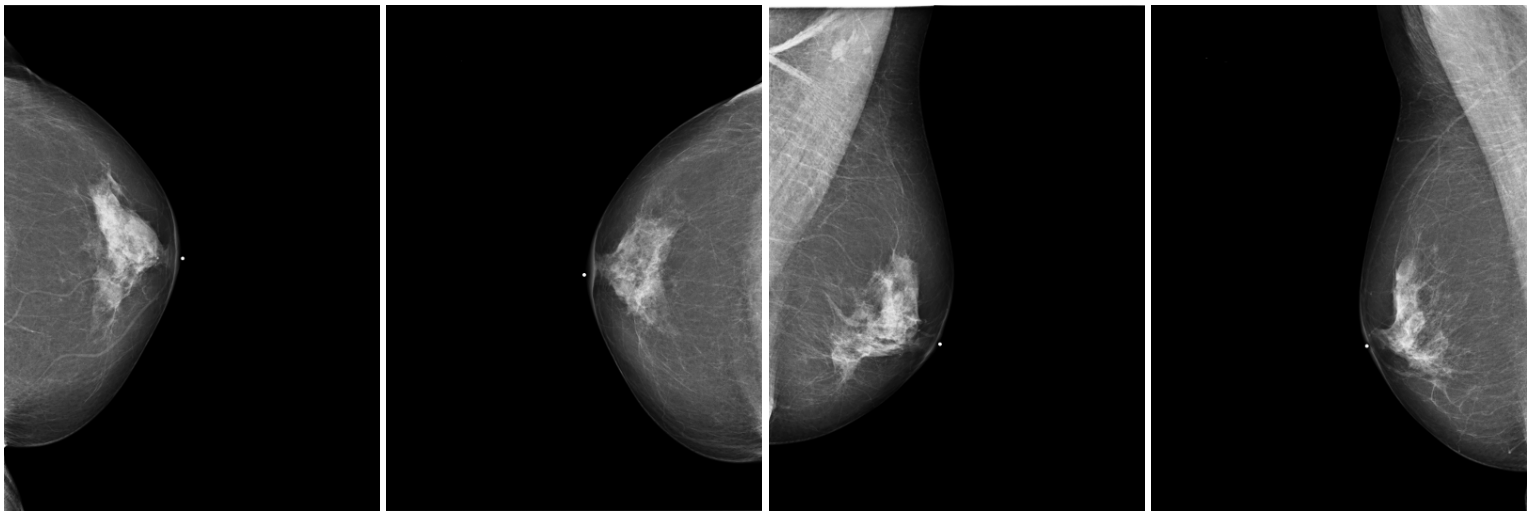
Breast cancer screening

- Data are X-ray images. Basically, an image recognition problem.
- Problem with a high impact on the society.
  - In 2015, 232k women in the US were diagnosed with breast cancer, 40k died.
- Results generalizable to other types of diseases.

Breast cancer screening

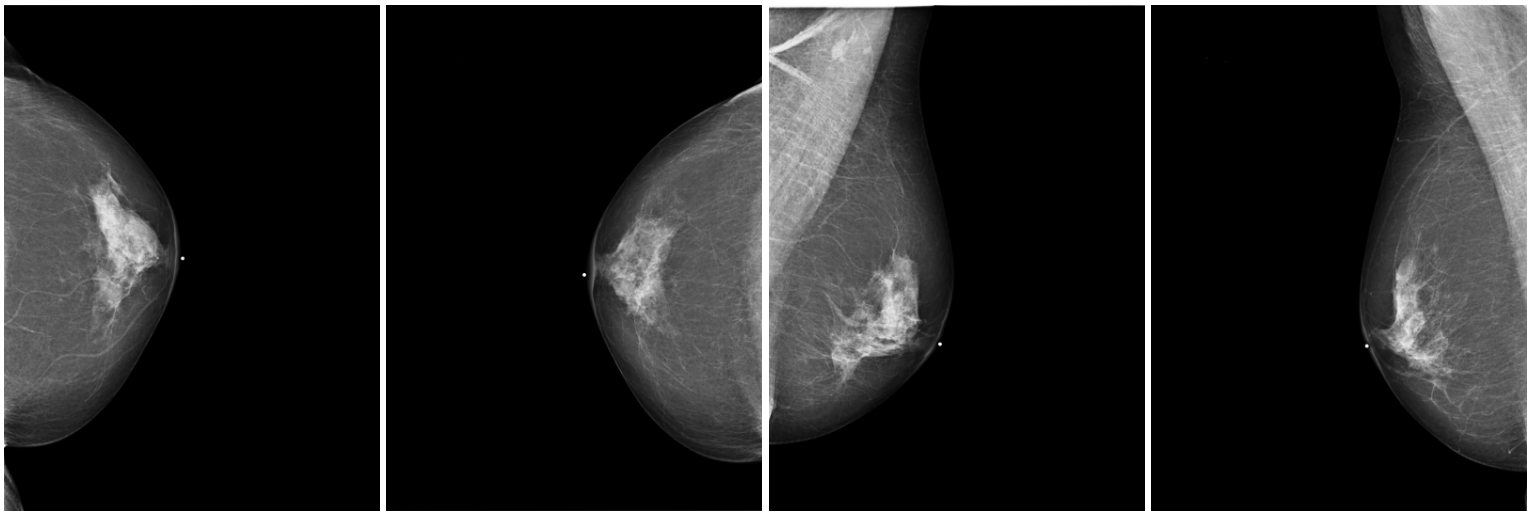
- Data are X-ray images. Basically, an image recognition problem.
- Problem with a high impact on the society.
  - In 2015, 232k women in the US were diagnosed with breast cancer, 40k died.
- Results generalizable to other types of diseases.
- Very little work done on end-to-end prediction. Fundamental scientific problems to be solved.

Breast cancer screening



L-CC	R-CC	L-MLO	R-MLO
(left cranial caudal)	(right cranial caudal)	(left mediolateral oblique)	(right mediolateral oblique)

Breast cancer screening



L-CC	R-CC	L-MLO	R-MLO
(left cranial caudal)	(right cranial caudal)	(left mediolateral oblique)	(right mediolateral oblique)

+ age, gender, cancer history, ethnicity, religion, ...

Breast cancer screening

We try to mimic prediction of a radiologist.

- Class 0: incomplete ($\approx 15\%$).
- Class 1: negative ($\approx 50\%$).
- Class 2: bening findings ($\approx 35\%$).

Breast cancer screening

We try to mimic prediction of a radiologist.

- Class 0: incomplete ($\approx 15\%$).
- Class 1: negative ($\approx 50\%$).
- Class 2: bening findings ($\approx 35\%$).

Radiologists call these classes BIRADS (short for Breast Imaging-Reporting and Data System).

Challenges (1)

You need a lot of data to do deep learning.

Challenges (1)

You need a lot of data to do deep learning.

Publicly available data sets contain about 1k images.

Challenges (1)

You need a lot of data to do deep learning.

Publicly available data sets contain about 1k images.

We build our own data set:

- 23k exams,
- 103k images.

Challenges (1)

You need a lot of data to do deep learning.

Publicly available data sets contain about 1k images.

We build our own data set:

- 23k exams,
- 103k images.

(And we are soon going to have $\times 10$ more).

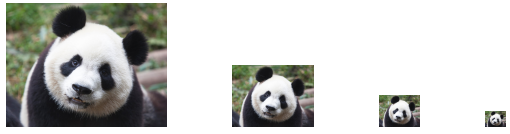
Challenges (2)

Challenges (2)

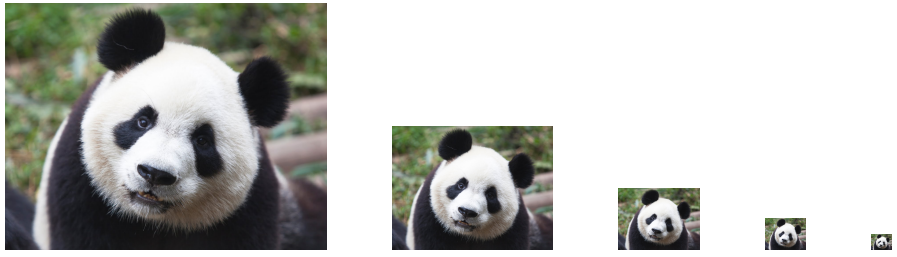
Challenges (2)



Challenges (2)



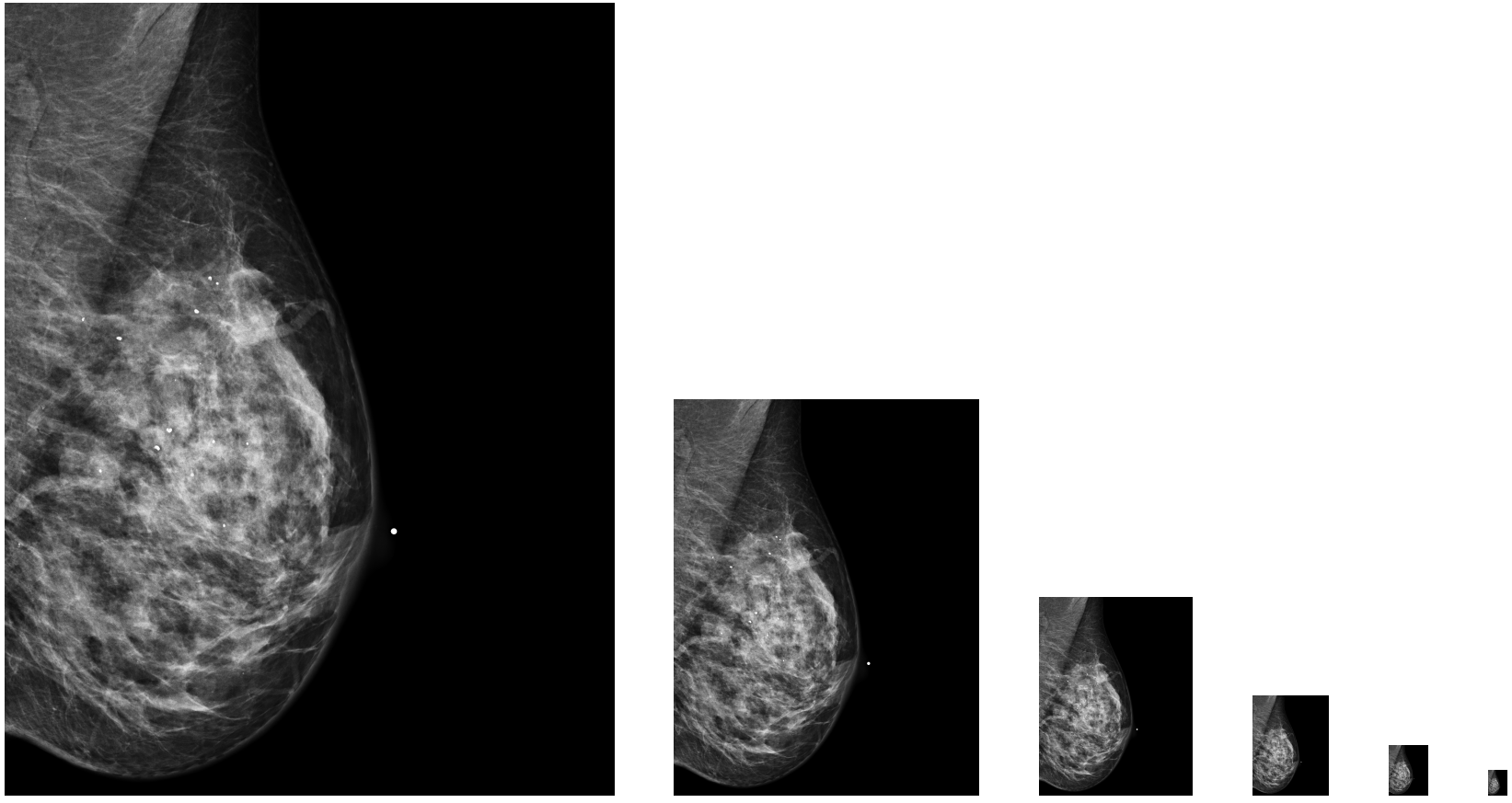
Challenges (2)



Challenges (2)

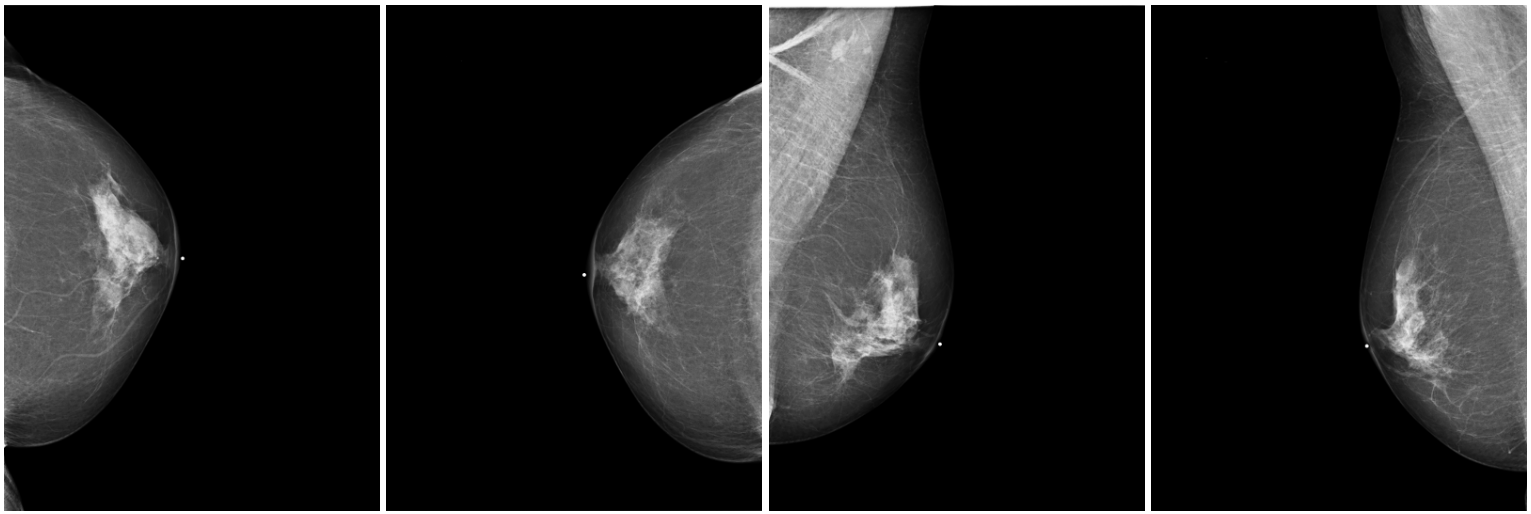
Challenges (2)

High resolution necessary - computational and engineering challenge.



Challenges (3)

Multi-view data. How to integrate information?



Our model

Classifier $p(y\|x)$			
Concatenation (256×4 dim)			
DCN	DCN	DCN	DCN
L-CC	R-CC	L-MLO	R-MLO

layer	**kernel size**	**stride**	**#maps**	**repetition**
global average pooling			256	
convolution	3×3	1×1	256	×3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	64	
convolution	3×3	1×1	64	× 2
convolution	3×3	2×2	64	
max pooling	3×3	3×3	32	
convolution	3×3	2×2	32	
input			1	

Our model

Classifier $p(y\|x)$			
Concatenation (256×4 dim)			
DCN	DCN	DCN	DCN
L-CC	R-CC	L-MLO	R-MLO

layer	kernel size	stride	#maps	repetition
global average pooling			256	
convolution	3×3	1×1	256	×3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	64	
convolution	3×3	1×1	64	× 2
convolution	3×3	2×2	64	
max pooling	3×3	3×3	32	
convolution	3×3	2×2	32	
input			1	

We use dropout, early stopping and data augmentation.

Our model

Classifier $p(y\|x)$			
Concatenation (256×4 dim)			
DCN	DCN	DCN	DCN
L-CC	R-CC	L-MLO	R-MLO

layer	kernel size	stride	#maps	repetition
global average pooling			256	
convolution	3×3	1×1	256	×3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	64	
convolution	3×3	1×1	64	× 2
convolution	3×3	2×2	64	
max pooling	3×3	3×3	32	
convolution	3×3	2×2	32	
input			1	

We use dropout, early stopping and data augmentation.
Columns share parameters.

Our model

Classifier $p(y\|x)$			
Concatenation (256×4 dim)			
DCN	DCN	DCN	DCN
L-CC	R-CC	L-MLO	R-MLO

layer	kernel size	stride	#maps	repetition
global average pooling			256	
convolution	3×3	1×1	256	×3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	64	
convolution	3×3	1×1	64	× 2
convolution	3×3	2×2	64	
max pooling	3×3	3×3	32	
convolution	3×3	2×2	32	
input			1	

We use dropout, early stopping and data augmentation.
Columns share parameters. We use Adam for learning.

Our model

Classifier $p(y\|x)$			
Concatenation (256×4 dim)			
DCN	DCN	DCN	DCN
L-CC	R-CC	L-MLO	R-MLO

layer	kernel size	stride	#maps	repetition
global average pooling			256	
convolution	3×3	1×1	256	×3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	128	
convolution	3×3	1×1	128	× 3
max pooling	2×2	2×2	64	
convolution	3×3	1×1	64	× 2
convolution	3×3	2×2	64	
max pooling	3×3	3×3	32	
convolution	3×3	2×2	32	
input			1	

We use dropout, early stopping and data augmentation.
Columns share parameters. We use Adam for learning.
A highly optimised implementation needs a week to train.

Evaluation metrics

- Per class AUC.

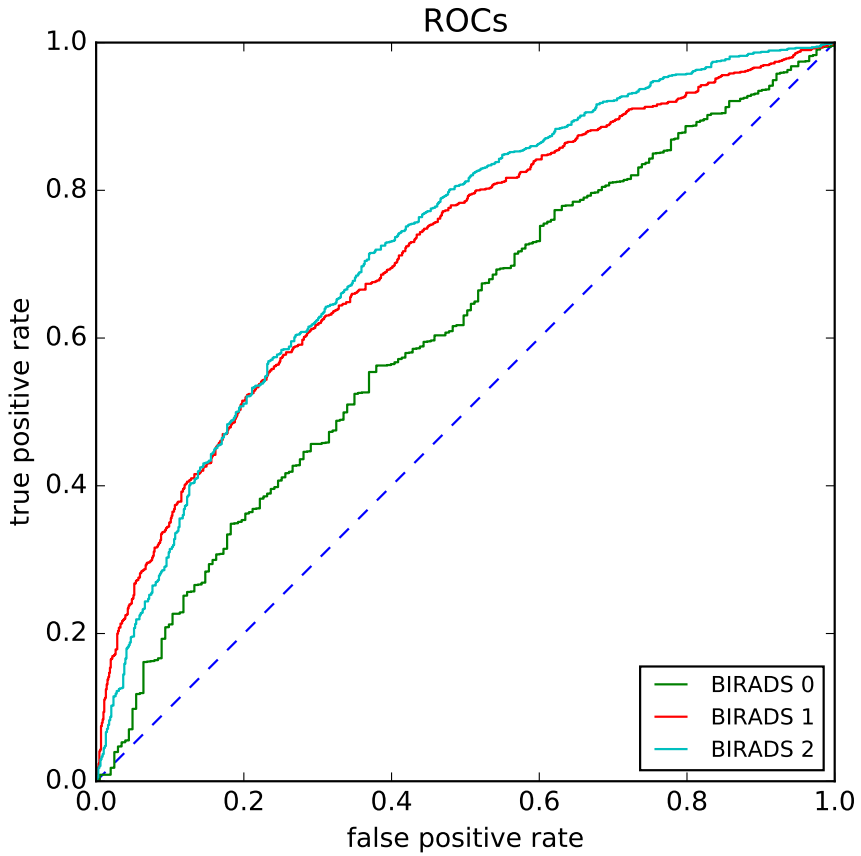
Evaluation metrics

- Per class AUC.
- Average AUC.

Evaluation metrics

- Per class AUC.
- Average AUC.
- Average AUC for 30% most confident examples.

Results

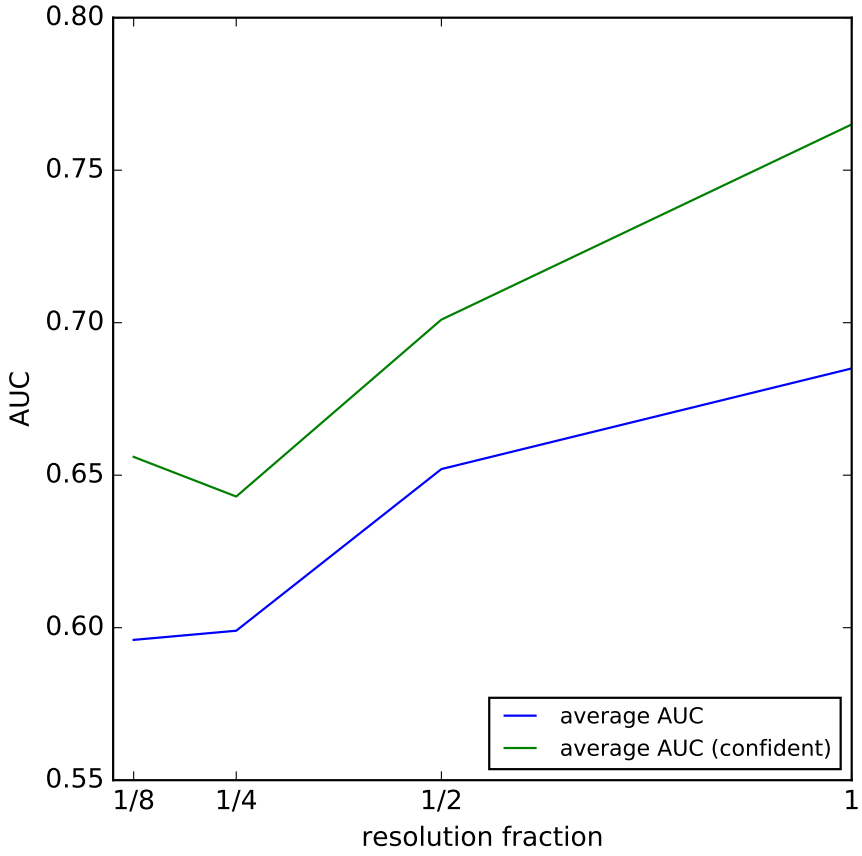


0 vs. others: 0.609
1 vs. others: 0.717
2 vs. others: 0.728

average: **0.685**
average (confident): **0.765**

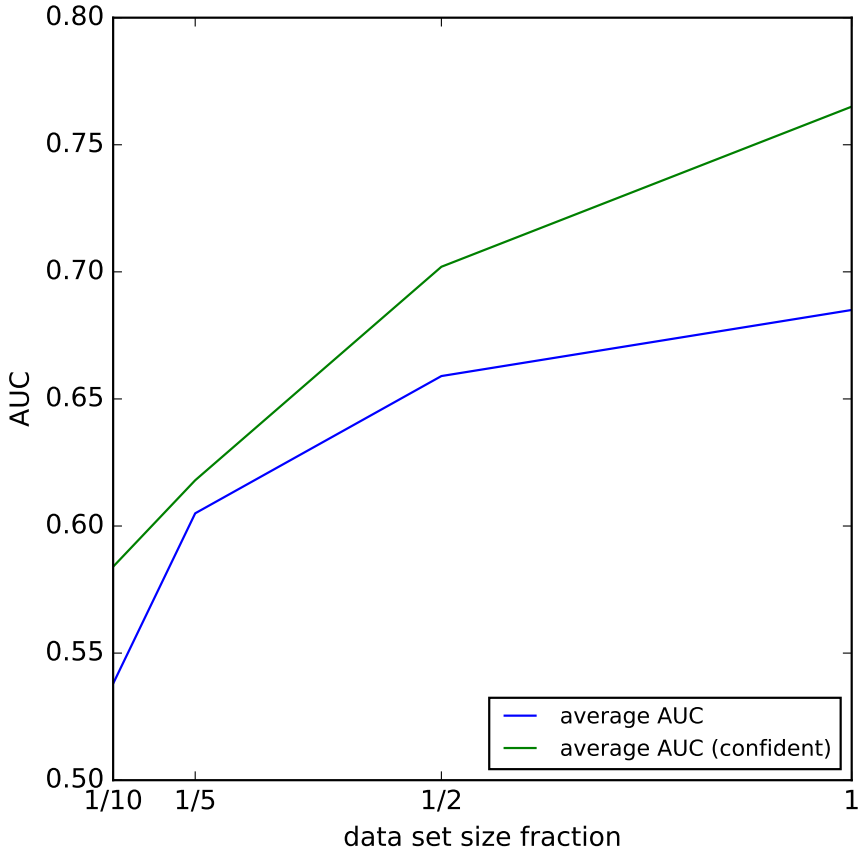
Impact of downscaling

Original image size - 2600×2000.



Impact of the data set size

Original data set size - 23k exams.

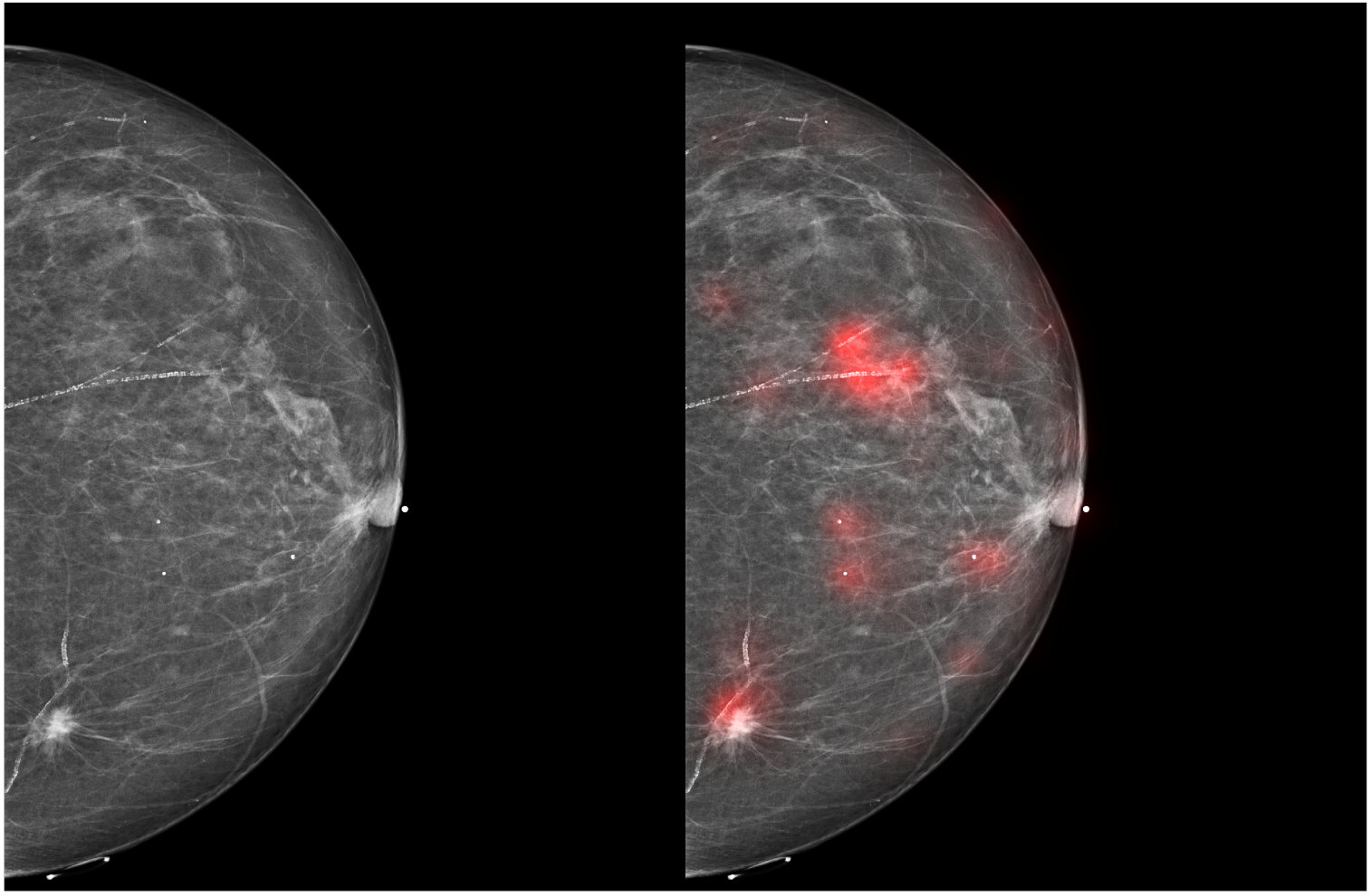


VISUALISATION

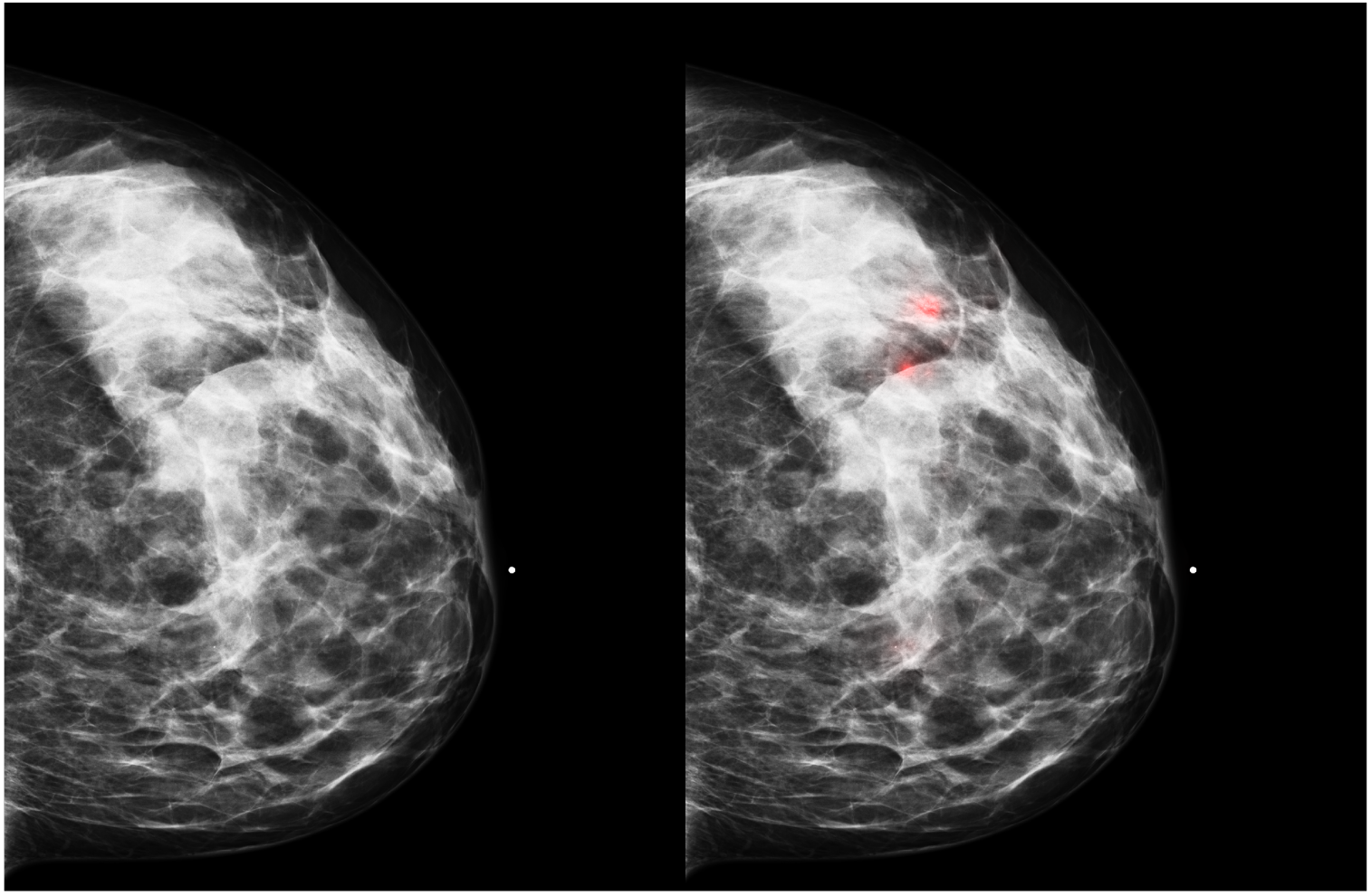
We visualise $\left| \frac{\partial H(y|\mathbf{x})}{\partial \mathbf{x}^v_{(i,j)}} \right|$,

where $H(y|\mathbf{x}) = -\sum_{y' \in \mathcal{C}} p(y'|\mathbf{x}) \log p(y'|\mathbf{x})$.

Visualisation



Visualisation



VISUALISATION

- Activations of the penultimate layer.

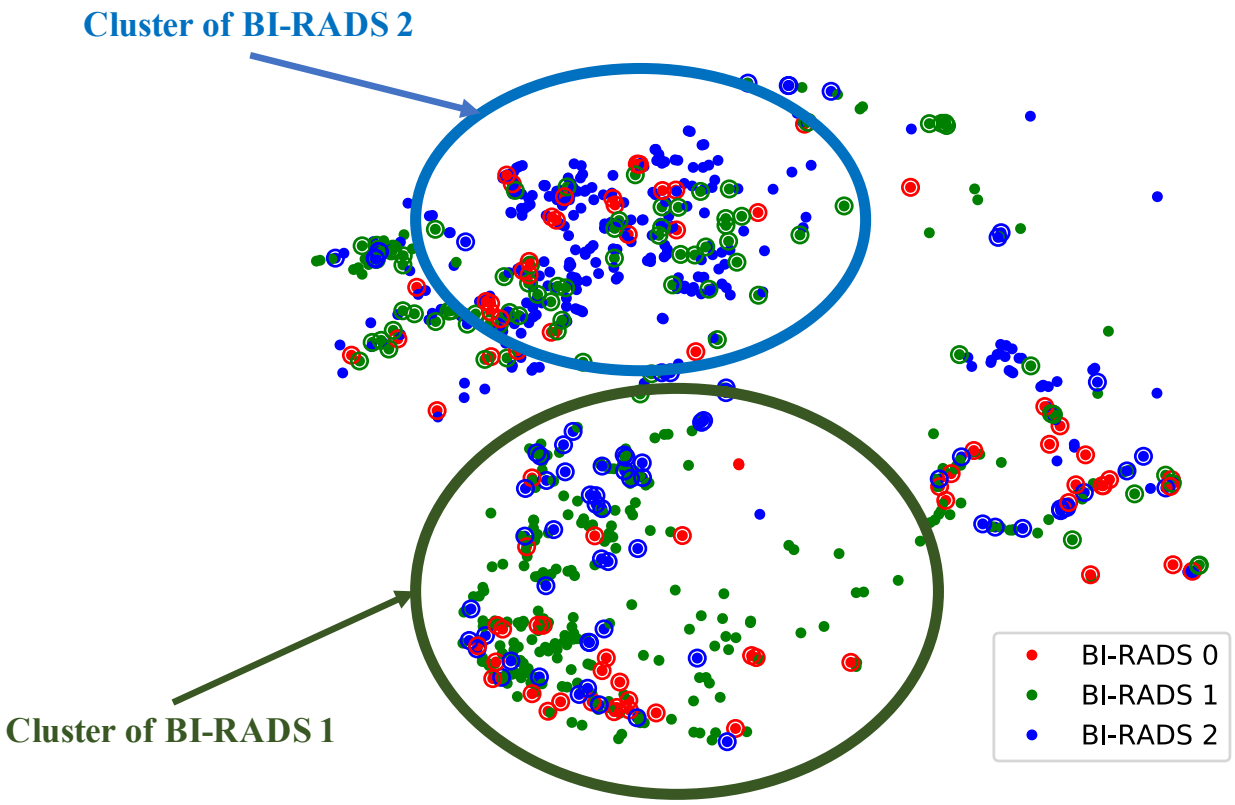
Visualisation

- Activations of the penultimate layer.
- t-SNE.

VISUALISATION

- Activations of the penultimate layer.
- t-SNE.
- Only confident examples shown.

VISUALISATION



Conclusions

Conclusions

- It is much harder to learn the “incomplete” (0) class than other classes.

Conclusions

- It is much harder to learn the “incomplete” (0) class than other classes.
- We need to use the full resolution.

Conclusions

- It is much harder to learn the “incomplete” (0) class than other classes.
- We need to use the full resolution.
- We need more data.

Ways to go forward

Ways to go forward

- Incorporating side information.

Ways to go forward

- Incorporating side information.
- Taking advantage of previous exams.

Ways to go forward

- Incorporating side information.
- Taking advantage of previous exams.
- Learning where to look.