Deep Learning in Radiology

Promise and Caveats

John R. Zech, M.D., M.A.
PGY-1 Prelim Medicine, CPMC

About Me: John Zech

Preliminary medicine intern, CPMC

Future radiology resident, Columbia

Studied machine learning at Columbia (M.A. Statistics)

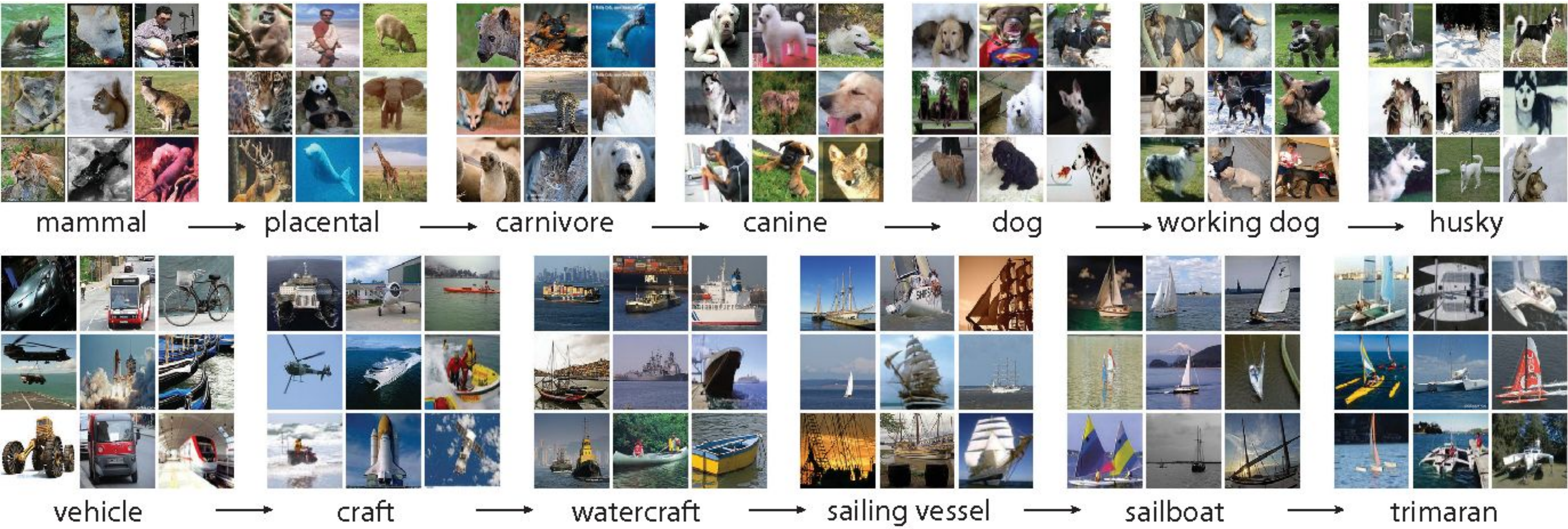
Prior to medicine: developed quantitative models in investment management

The rise of the machines

The rise of the machines: digit recognition

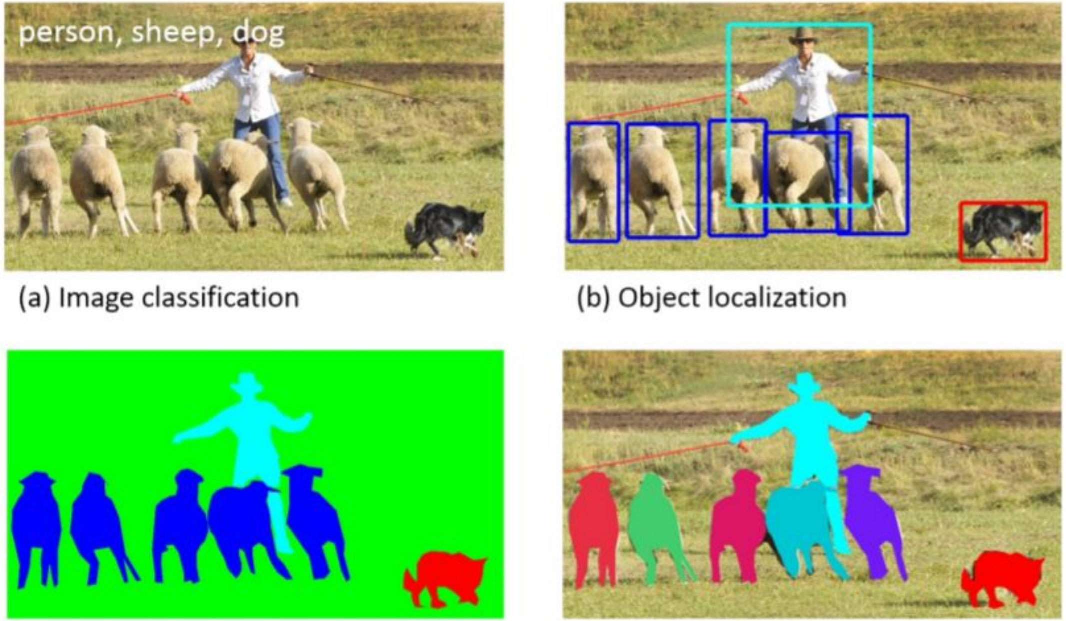
LeCun et al., 1998

The rise of the machines: object recognition



Deng et al., 2009, Russakovsky et al., 2015

Segmentation CNNs



(a) Image classification

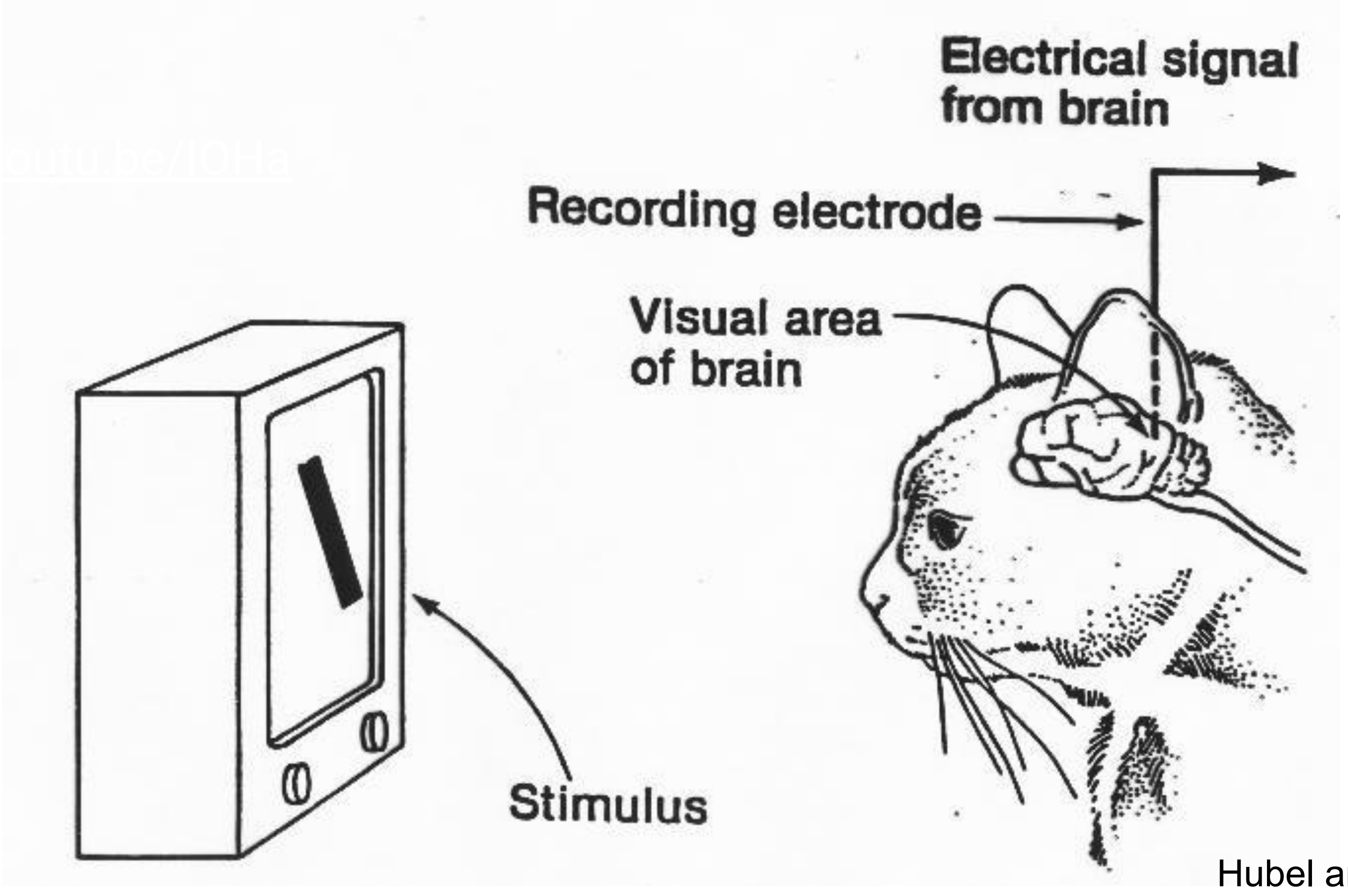
(b) Object localization

(c) Semantic segmentation

(d) This work

https://arxiv.org/pdf/1405.0312.pdf

Neural networks: biologically inspired

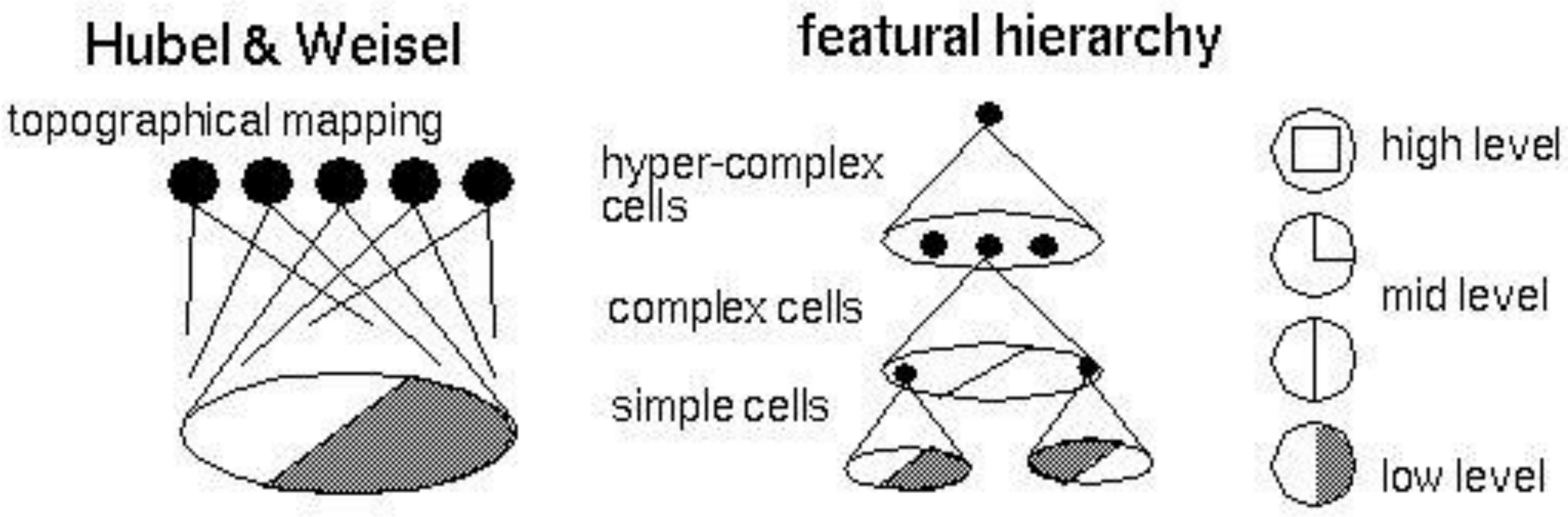


Hubel and Wiesel, 1962

Neural networks: biologically inspired

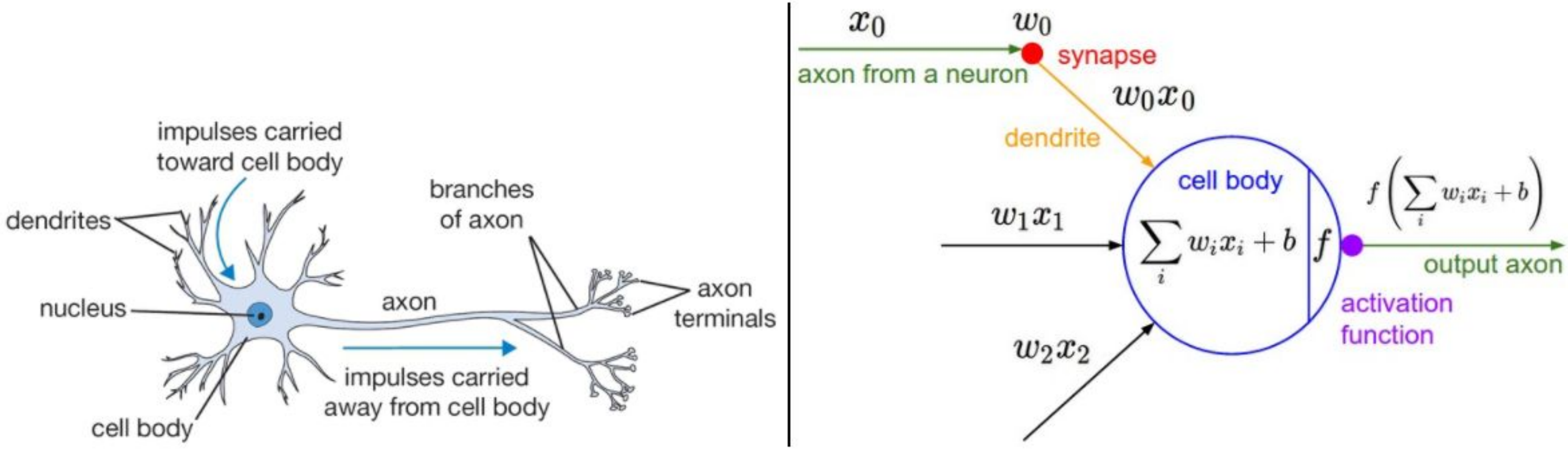
Hubel and Wiesel, 1962 - https://youtu.be/IOHayh06LJ4

Neural networks: biologically inspired



Hubel and Weisel, 1962

Neural networks: biologically inspired

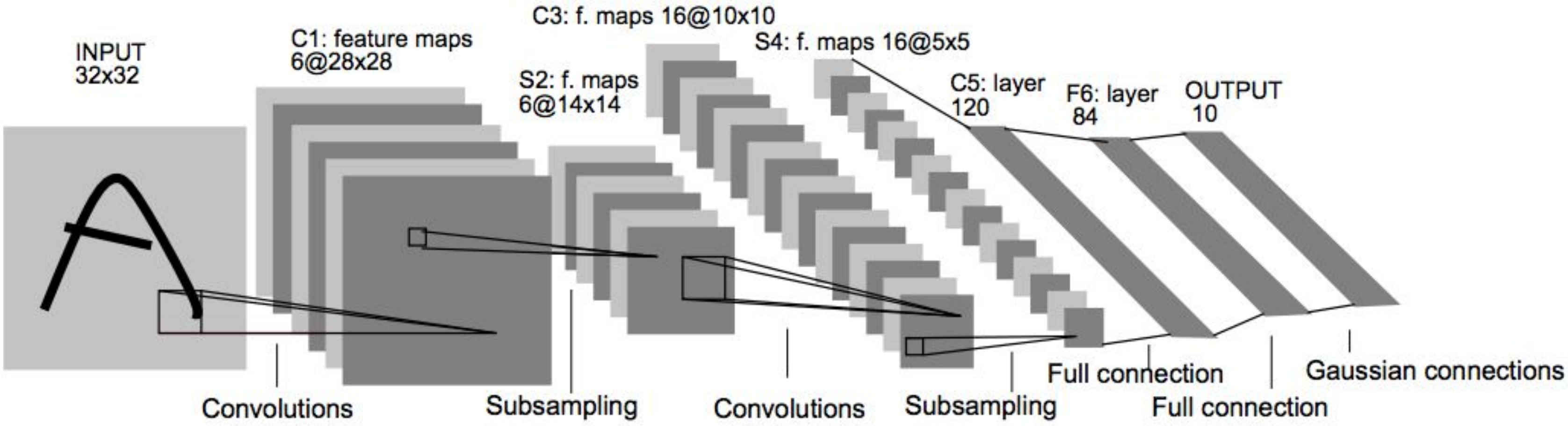


A cartoon drawing of a biological neuron (left) and its mathematical model (right).

http://cs231n.github.io/

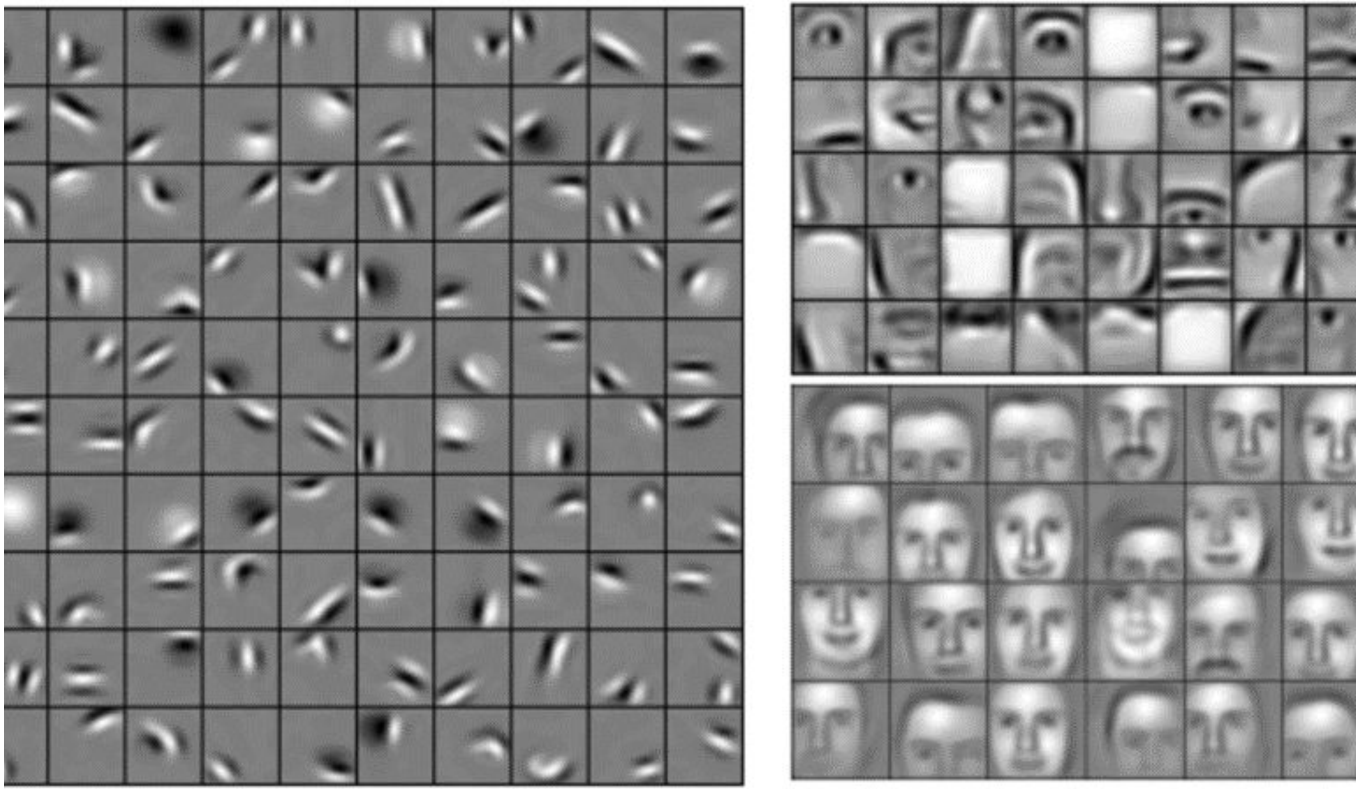
Classification CNNs

Maps image to a single classification



https://medium.com/@pechyonkin/key-deep-learning-architectures-lenet-5-6fc3c59e6f4

Learned features are *sometimes* interpretable



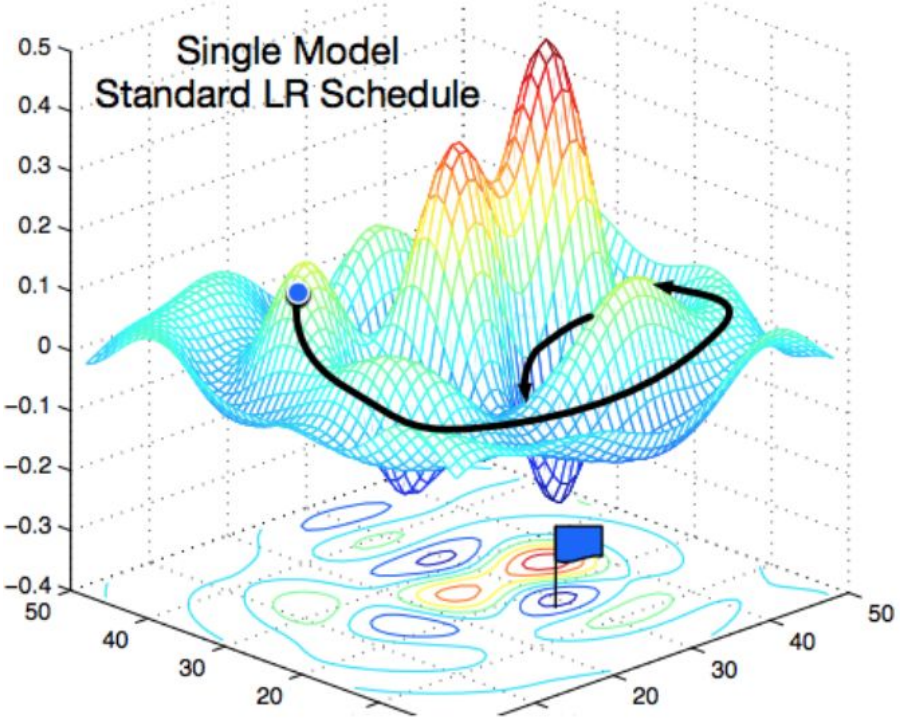
Mid-level

High-level

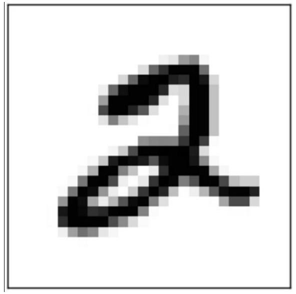
Lower level

Lee et al., 2009

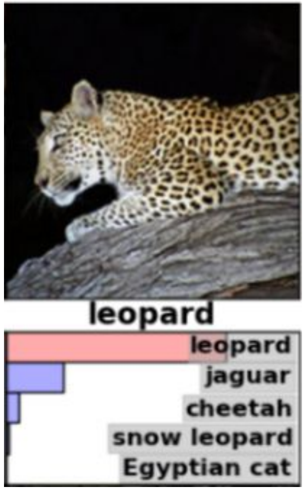
CNNs are trained in an iterative process using stochastic gradient descent



CNNs have gotten more complex over 20 years



MNIST: 32 x 32 pixels



ImageNet: varies, 224 x 224 - 299 x 299

CNNs have gotten more complex over 20 years

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)

CNNs have gotten more complex over 20 years

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)

IT'S NOT ABOUT THE ALGORITHM

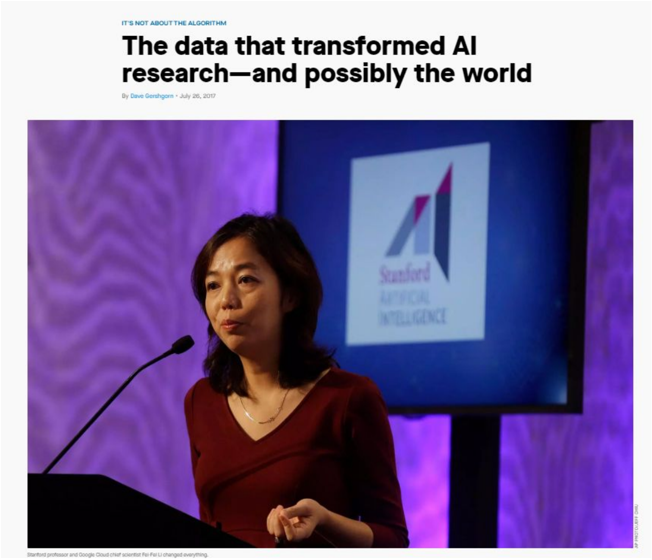
The data that transformed AI research—and possibly the world

By Dave Gershgorn • July 26, 2017

Stanford professor and Google Cloud chief scientist Fei-Fei Li changed everything.

CNNs have gotten more complex over 20 years

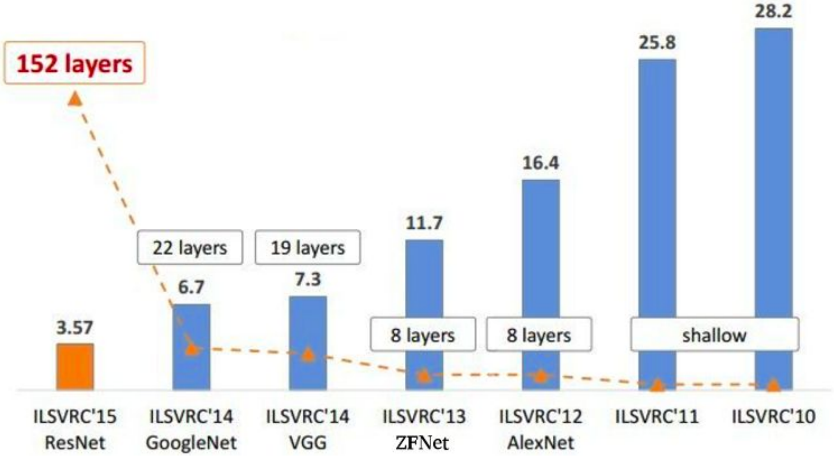
IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)


IT'S NOT ABOUT THE ALGORITHM

The data that transformed AI research—and possibly the world

By Dave Gershgorn • July 26, 2017

Stanford professor and Google Cloud chief scientist Fei-Fei Li changed everything.

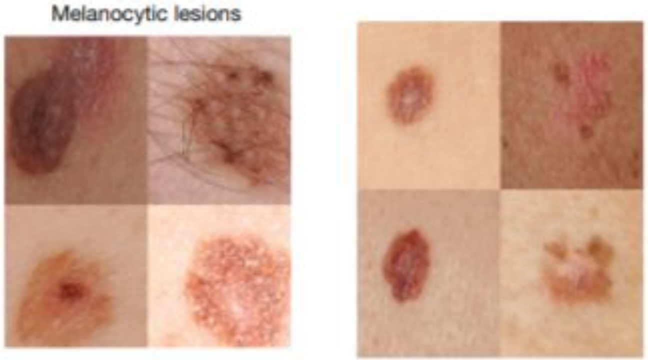


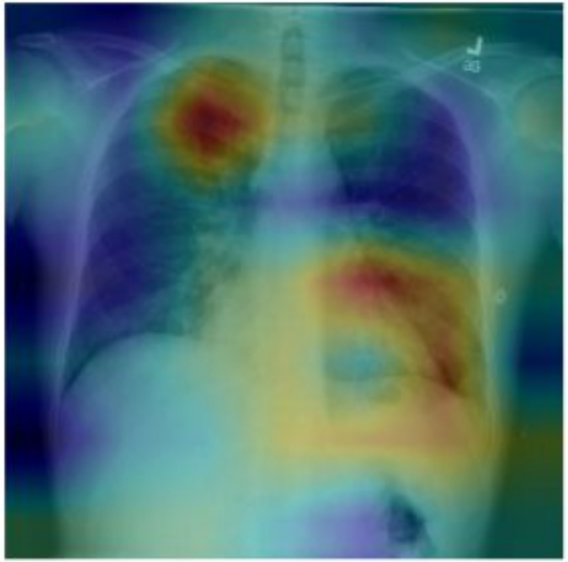
https://software.intel.com/en-us/articles/hands-on-ai-part-16-modern-deep-neural-network-architectures-for-image-classification

CNNs are challenging to train, but...

You can start with pre-trained model and ‘fine-tune’ to your problem

The rise of the machines: two case-studies in human-level clinical prediction





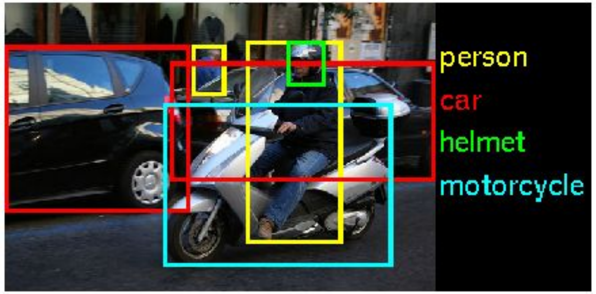
Esteva et al. (2017)

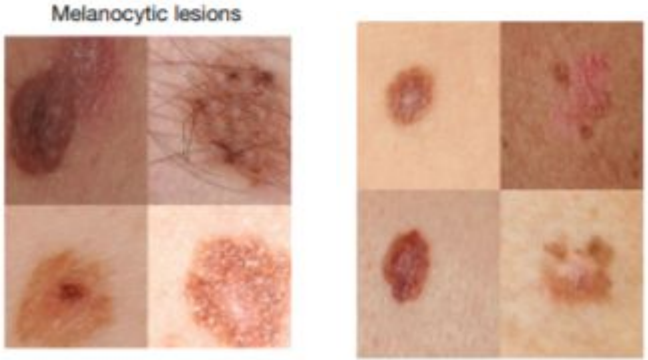
Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

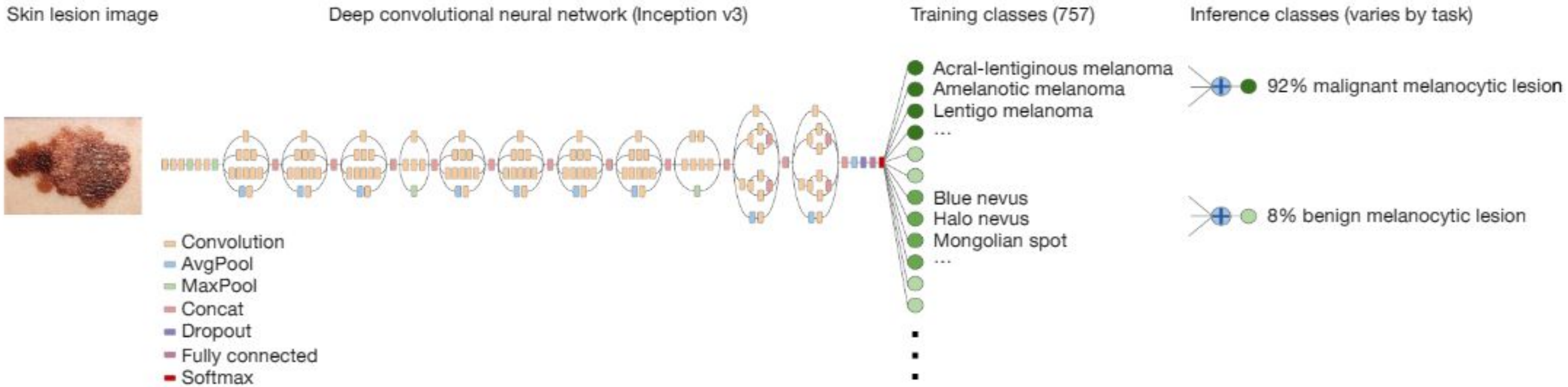
Esteva et al. (2017)

- Inception v3 model (299 x 299) pre-trained in another domain (ImageNet)
- Fine-tuned CNN with 129,450 clinical images



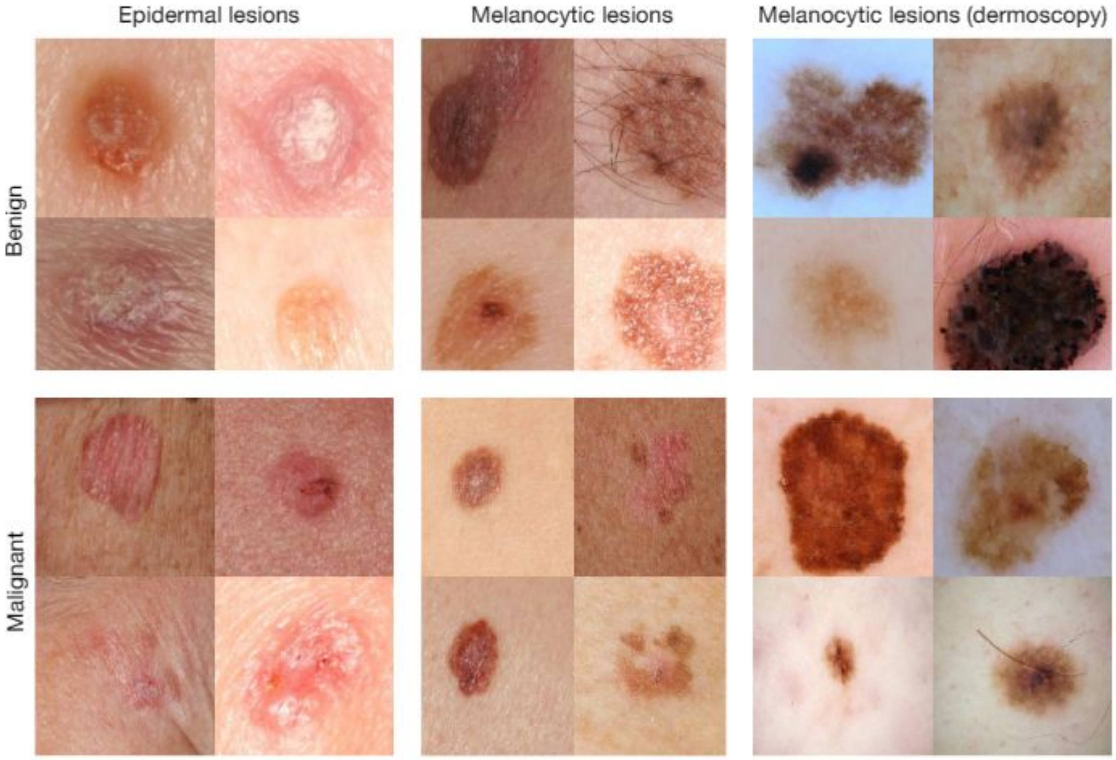


Esteva et al. (2017)



Esteva et al. (2017)

All comparison happened on 1,942 held-out biopsy-proven images: **strong ground truth**

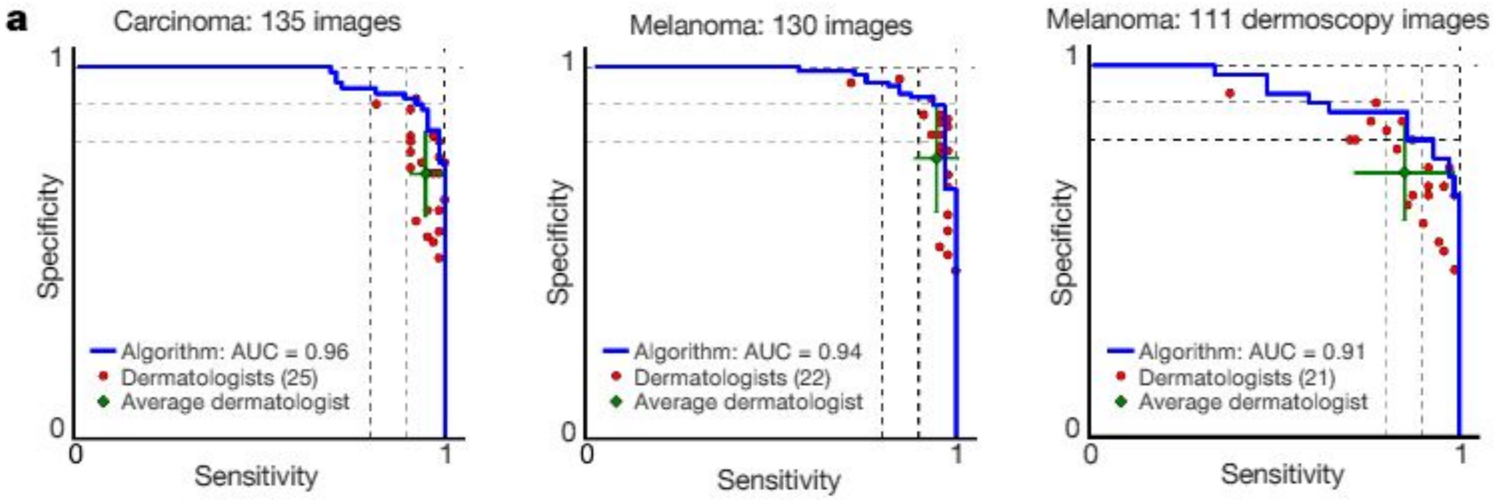


Esteva et al. (2017)

Compare to 21 dermatologists on:

1. Keratinocyte carcinomas vs benign seborrheic keratoses (most common skin cancer)
2. Malignant melanomas versus benign nevi (most deadly skin cancer)

Esteva et al. (2017)



What worked well in Esteva et al. (2017) :

- Image resolution not a limitation
- Clinical information outside the image may have limited value
- Strong ground truth comparison: biopsy results

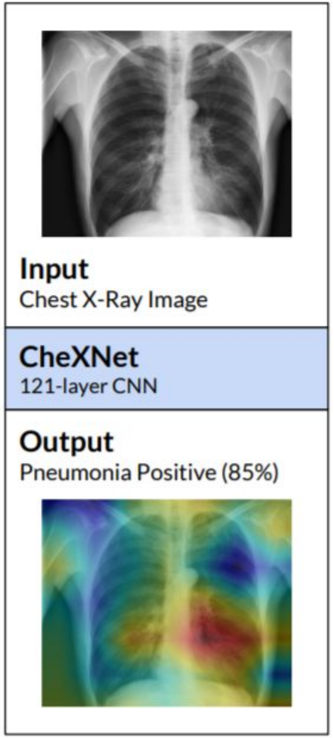
Rajpurkar et al. (2017)

CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar[*1] Jeremy Irvin[*1] Kaylie Zhu[1] Brandon Yang[1] Hershel Mehta[1]
Tony Duan[1] Daisy Ding[1] Aarti Bagul[1] Robyn L. Ball[2] Curtis Langlotz[3] Katie Shpanskaya[3]
Matthew P. Lungren[3] Andrew Y. Ng[1]

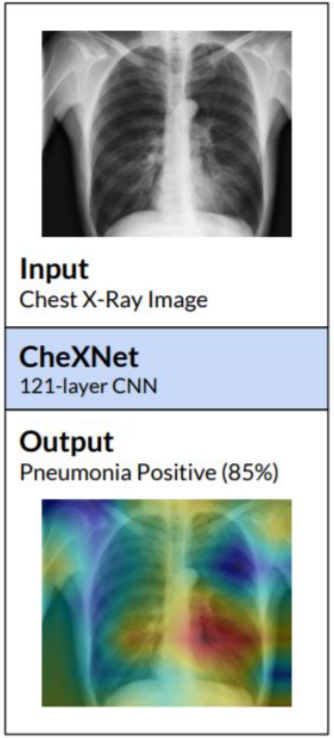
Rajpurkar et al. (2017)

- Pre-trained DenseNet-121
  - 224 x 224 pixels



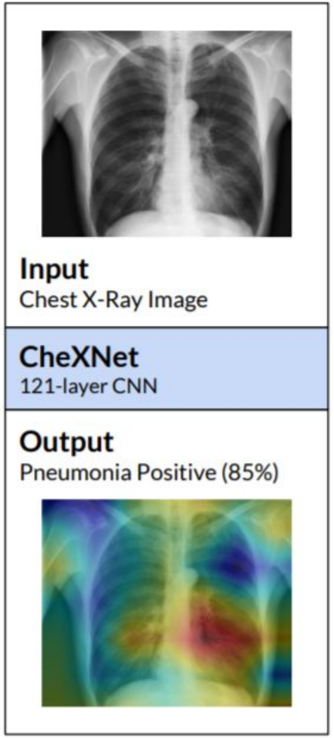
Rajpurkar et al. (2017)

- Pre-trained DenseNet-121
  - 224 x 224 pixels
- 112,120 NIH chest x-rays
  - 70% train, 10% tune, 20% test



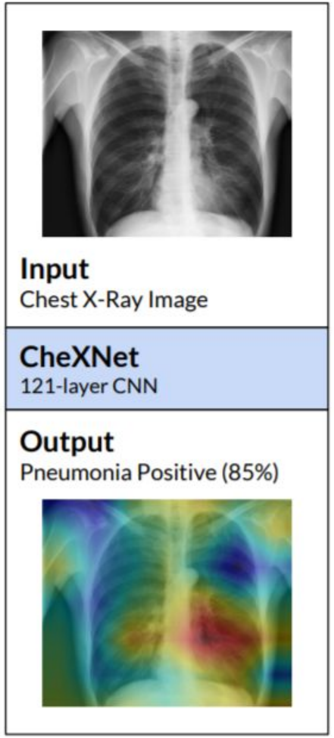
Rajpurkar et al. (2017)

- Pre-trained DenseNet-121
  - 224 x 224 pixels
- 112,120 NIH chest x-rays
  - 70% train, 10% tune, 20% test
- 14 diagnoses, including pneumonia



Rajpurkar et al. (2017)

- 14 diagnoses, including pneumonia
- AUC for pneumonia: 0.7680



Rajpurkar et al. (2017)

- Human comparison: special 420 x-ray test set, labeled by 4 Stanford radiologists.

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

$$F_1 = \frac{2}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

https://en.wikipedia.org/wiki/F1_score

Rajpurkar et al. (20

- Human comparison: special 420 x-ray test set, labeled by 4 Stanford radiologists.

	F1 Score (95% CI)
Radic	0.383 (0.309, 0.453)
Ra 2	0.356 (0.282, 0.428)
ist 3	0.365 (0.291, 0.435)
logist 4	0.442 (0.390, 0.492)
iologist Avg.	0.387 (0.330, 0.442)
XNet	0.435 (0.387, 0.481)

https://en.wikipedia.org _score

$$F_1 = \frac{2}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Human-level performance?

- AUC of Rajpurkar et al. (2017): for pneumonia 0.7680
    - By comparison, AUC of Esteva et al. (2017): 0.91-0.96
- Why did the Rajpurkar et al. (2017) compare using 4 radiologists and F1 score?
    - Low radiologist agreement

Reproduce-CheXNet: Zech (2018)

GitHub, Inc. [US] | https://github.com/jrzech/reproduce-chexnet

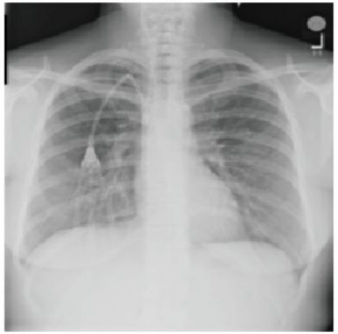
README.md

reproduce-chexnet

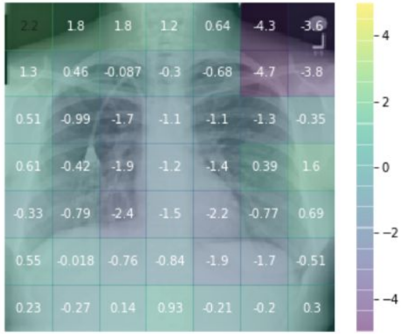
launch binder

Provides Python code to reproduce model training, predictions, and heatmaps from the CheXNet paper that predicted 14 common diagnoses using convolutional neural networks in over 100,000 NIH chest x-rays.

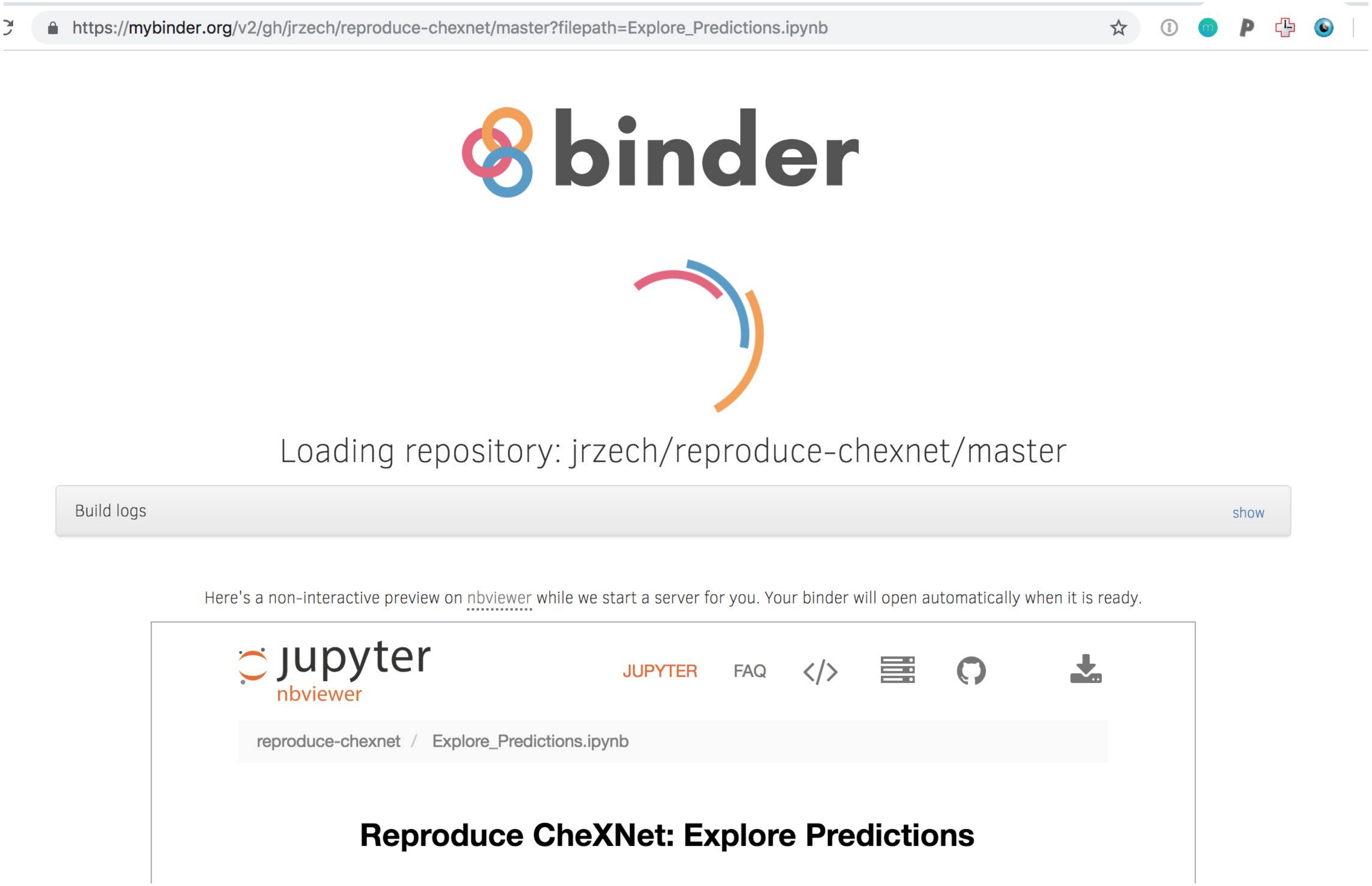
00004075_000.png



P(Pneumonia)=0.007



https://github.com/jrzech/reproduce-chexnet


https://mybinder.org/v2/gh/jrzech/reproduce-chexnet/master?filepath=Explore_Predictions.ipynb
binder
Loading repository: jrzech/reproduce-chexnet/master
Build logs
show
Here's a non-interactive preview on nbviewer while we start a server for you. Your binder will open automatically when it is ready.
jupyter
nbviewer
JUPYTER
FAQ
reproduce-chexnet / Explore_Predictions.ipynb
Reproduce CheXNet: Explore Predictions

Reproduce-CheXNet: Zech (2018)

```
def train_model(
        model,
        criterion,
        optimizer,
        LR,
        num_epochs,
        dataloaders,
        dataset_sizes,
        weight_decay):
    """
    Fine tunes torchvision model to NIH CXR data.

    Args:
        model: torchvision model to be finetuned (densenet-121 in this case)
        criterion: loss criterion (binary cross entropy loss, BCELoss)
        optimizer: optimizer to use in training (SGD)
        LR: learning rate
        num_epochs: continue training up to this many epochs
        dataloaders: pytorch train and val dataloaders
        dataset_sizes: length of train and val datasets
        weight_decay: weight decay parameter we use in SGD with momentum
    Returns:
        model: trained torchvision model
        best_epoch: epoch on which best model val loss was obtained

    """
    since = time.time()

    start_epoch = 1
```

	retrained auc	chexnet auc
label		
Atelectasis	0.8161	0.8094
Cardiomegaly	0.9105	0.9248
Consolidation	0.8008	0.7901
Edema	0.8979	0.8878
Effusion	0.8839	0.8638
Emphysema	0.9227	0.9371
Fibrosis	0.8293	0.8047
Hernia	0.9010	0.9164
Infiltration	0.7077	0.7345
Mass	0.8308	0.8676
Nodule	0.7748	0.7802
Pleural_Thickening	0.7860	0.8062
Pneumonia	0.7651	0.7680
Pneumothorax	0.8739	0.8887

Reproduce-CheXNet: Zech (2018)

```
def train_model(
        model,
        criterion,
        optimizer,
        LR,
        num_epochs,
        dataloaders,
        dataset_sizes,
        weight_decay):
    """
    Fine tunes torchvision model to NIH CXR data.

    Args:
        model: torchvision model to be finetuned (densenet-121 in this case)
        criterion: loss criterion (binary cross entropy loss, BCELoss)
        optimizer: optimizer to use in training (SGD)
        LR: learning rate
        num_epochs: continue training up to this many epochs
        dataloaders: pytorch train and val dataloaders
        dataset_sizes: length of train and val datasets
        weight_decay: weight decay parameter we use in SGD with momentum
    Returns:
        model: trained torchvision model
        best_epoch: epoch on which best model val loss was obtained

    """
    since = time.time()

    start_epoch = 1
```

Similar to Rajpurkar et al. (2017): 0.7680 vs. 0.7651

	retrained auc	chexnet auc
label		
Atelectasis	0.8161	0.8094
Cardiomegaly	0.9105	0.9248
Consolidation	0.8008	0.7901
Edema	0.8979	0.8878
Effusion	0.8839	0.8638
Emphysema	0.9227	0.9371
Fibrosis	0.8293	0.8047
Hernia	0.9010	0.9164
Infiltration	0.7077	0.7345
Mass	0.8308	0.8676
Nodule	0.7748	0.7802
Pleural_Thickening	0.7860	0.8062
Pneumonia	0.7651	0.7680
Pneumothorax	0.8739	0.8887

Will CheXNet generalize?

Confounders in Radiology: Zech et al. 2018

Confounding variables can degrade generalization performance of radiological deep learning models

John R. Zech[1*], Marcus A. Badgeley[2*], Manway Liu[2], Anthony B. Costa[3], Joseph J. Titano[4], Eric K. Oermann[3]

1 Department of Medicine, California Pacific Medical Center, San Francisco, CA 94115
`jrz2111@columbia.edu`
2 Verily Life Sciences, 269 E Grand Ave, South San Francisco, CA 94080
`marcus.badgeley@icahn.mssm.edu, manwayl@verily.com`
3 Department of Neurological Surgery, Icahn School of Medicine, New York, NY 10029
`anthony.costa@mountsinai.org, eric.oermann@mountsinai.org`
4 Department of Radiology, Icahn School of Medicine, New York, NY 10029
`joseph.titano@mountsinai.org`

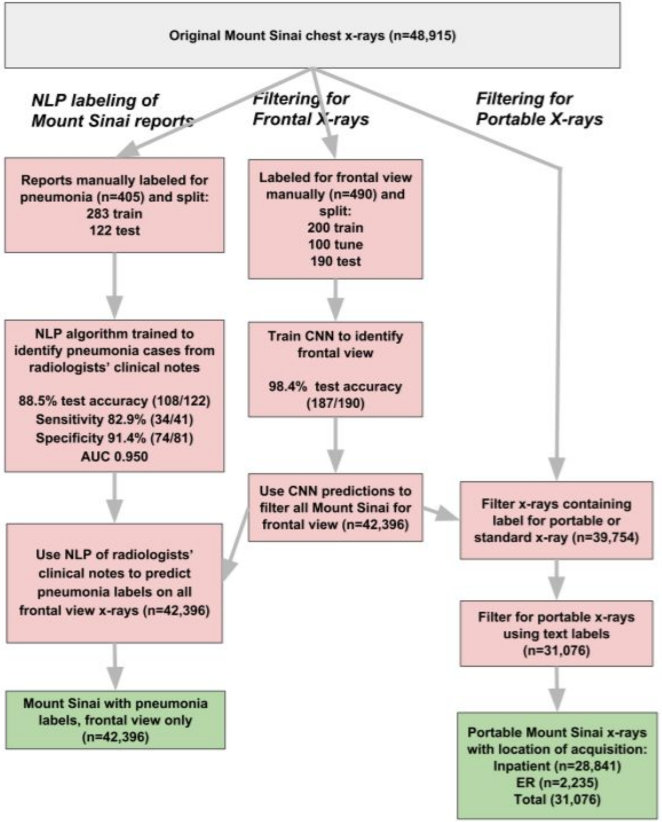
* These authors contributed equally to this work

Confounders in Radiology: Zech et al. 2018

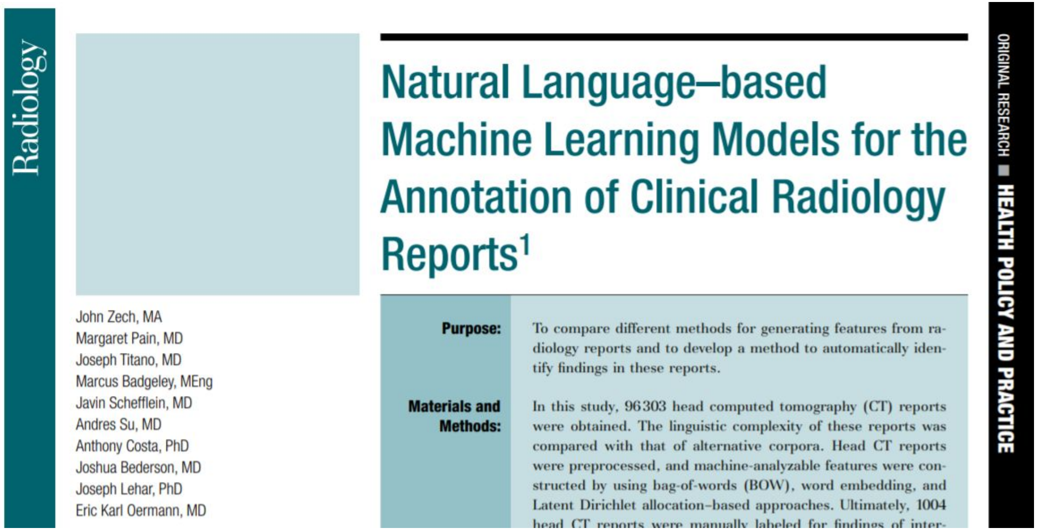
- How well does CheXNet generalize?
- Trained CheXNet using data from
  - NIH
  - Mount Sinai
  - Indiana University
- Trained / tested using different combinations of data sources

Building the Mount Sinai dataset: Zech et al. 2018

- Exported and preprocessed 48,915 DICOM files from Mt. Sinai PACS
- **Used NLP to automatically infer labels**



Building the Mount Sinai dataset: Zech et al. 2018



Radiology

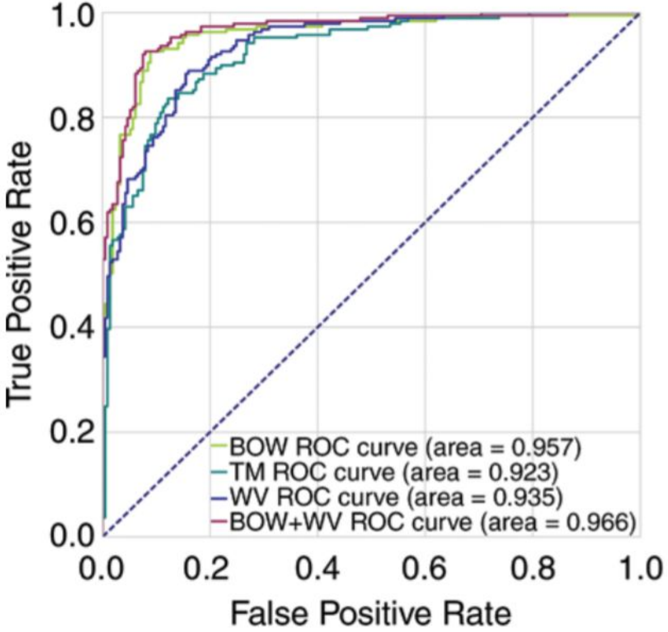
ORIGINAL RESEARCH ■ HEALTH POLICY AND PRACTICE

Natural Language–based Machine Learning Models for the Annotation of Clinical Radiology Reports[1]

John Zech, MA
Margaret Pain, MD
Joseph Titano, MD
Marcus Badgeley, MEng
Javin Schefflein, MD
Andres Su, MD
Anthony Costa, PhD
Joshua Bederson, MD
Joseph Lehar, PhD
Eric Karl Oermann, MD

Purpose: To compare different methods for generating features from radiology reports and to develop a method to automatically identify findings in these reports.

Materials and Methods: In this study, 96303 head computed tomography (CT) reports were obtained. The linguistic complexity of these reports was compared with that of alternative corpora. Head CT reports were preprocessed, and machine-analyzable features were constructed by using bag-of-words (BOW), word embedding, and Latent Dirichlet allocation–based approaches. Ultimately, 1004



Building the Mount Sinai dataset: Zech et al. 2018

Exam Number: 12345678 Report Status: Final
Type: Chest 2 Views
Date/Time: 01/01/2014 10:30
Exam Code: XRCH2
Ordering Provider: Wayne, John Michael MD

HISTORY:
- Cough and Fever

REPORT Frontal and lateral views of the chest.

COMPARISON: None

FINDINGS:
Lines/tubes: None.

Lungs: The lungs are well inflated and clear. There is no evidence of pneumonia or pulmonary edema.

Pleura: There is no pleural effusion or pneumothorax.

Heart and mediastinum: The cardiomediastinal silhouette is normal.

Bones: The visualized skeleton is normal.

IMPRESSION:
Clear lungs without evidence of pneumonia.

RECOMMENDATION:
None.

PROVIDERS:
Doe, Jane Lynn MD

SIGNATURES:
Doe, Jane Lynn MD

If you have questions or concerns regarding this report, feel free to contact us by phone at 555-555-5555, or by e-mail at contact@aplusradiology.com

Building the Mount Sinai dataset: Zech et al. 2018

Exam Number: 12345678 Report Status: Final
Type: Chest 2 Views
Date/Time: 01/01/2014 10:30
Exam Code: XRCH2
Ordering Provider: Wayne, John Michael MD

HISTORY:
- Cough and Fever

REPORT Frontal and lateral views of the chest.

COMPARISON: None

FINDINGS:
Lines/tubes: None.

Lungs: The lungs are well inflated and clear. There is no evidence of pneumonia or pulmonary edema.

Pleura: There is no pleural effusion or pneumothorax.

Heart and mediastinum: The cardiomediastinal silhouette is normal.

Bones: The visualized skeleton is normal.

IMPRESSION:
Clear lungs without evidence of pneumonia.

RECOMMENDATION:
None.

PROVIDERS:
Doe, Jane Lynn MD

SIGNATURES:
Doe, Jane Lynn MD

If you have questions or concerns regarding this report, feel free to contact us by phone at 555-555-5555, or by e-mail at contact@aplusradiology.com

"Without evidence of pneumonia"

Building the Mount Sinai dataset: Zech et al. 2018

- These are imperfect labels
  - ~90% sensitivity, specificity

Building the Mount Sinai dataset: Zech et al. 2018

- These are imperfect labels
  - ~90% sensitivity, specificity
- What could introduce biases into these labels?

Building the Mount Sinai dataset: Zech et al. 2018

- These are imperfect labels
    - ~90% sensitivity, specificity
- What could introduce biases into these labels?
    - Radiologist thresholds for calling pathology
    - Institutional templates
    - Clinical scenario (i.e. ICU films for line placement)

Confounders in Radiology: Zech et al. 2018

Table 3. Internal and external pneumonia screening performance for all train - tune and test hospital system combinations.

Train - Tune Site	Comparison Type*	Test Site (Images)	AUC (95% C.I.)	Acc.	Sens.	Spec.	PPV	NPV
NIH	Internal	NIH (N=22,062)	0.750 (0.721-0.778)	0.255	0.951	0.247	0.015	0.998
	External	MSH (N=8,388)	0.695 (0.683-0.706)	0.476	0.950	0.212	0.401	0.884
	External	IU (N=3,807)	0.725 (0.644-0.807)	0.190	0.974	0.182	0.012	0.999
	Superset	MSH + NIH (N=30,450)	0.773 (0.766-0.780)	0.462	0.950	0.403	0.160	0.985
	Superset	MSH + NIH + IU (N=34,257)	0.787 (0.780-0.793)	0.470	0.950	0.418	0.148	0.987
MSH	Internal	MSH (N=8,388)	0.802 (0.793-0.812)	0.617	0.950	0.432	0.482	0.940
	External	NIH (N=22,062)	0.717 (0.687-0.746)	0.184	0.951	0.175	0.014	0.997
	External	IU (N=3,807)	0.756 (0.674-0.838)	0.099	0.974	0.090	0.011	0.997
	Superset	MSH + NIH (N=30,450)	0.862 (0.856-0.868)	0.562	0.950	0.516	0.190	0.989
	Superset	MSH + NIH + IU (N=34,257)	0.871 (0.865-0.877)	0.577	0.950	0.537	0.180	0.990
MSH + NIH	Internal	MSH + NIH (N=30,450)	0.931 (0.927-0.936)	0.732	0.950	0.706	0.279	0.992
	Subset	NIH (N=22,062)	0.733 (0.703-0.762)	0.243	0.951	0.234	0.015	0.997
	Subset	MSH (N=8,388)	0.805 (0.796-0.814)	0.630	0.950	0.451	0.491	0.942
	External	IU (N=3,807)	0.815 (0.745-0.885)	0.238	0.974	0.230	0.013	0.999
	Superset	MSH + NIH + IU (N=34,257)	0.934 (0.929-0.938)	0.732	0.950	0.709	0.258	0.993

*Superset= a test dataset containing data from the same distribution (hospital system) as the training data as well as external data. Subset = a test dataset containing data from fewer distributions (hospital systems) then the training data.

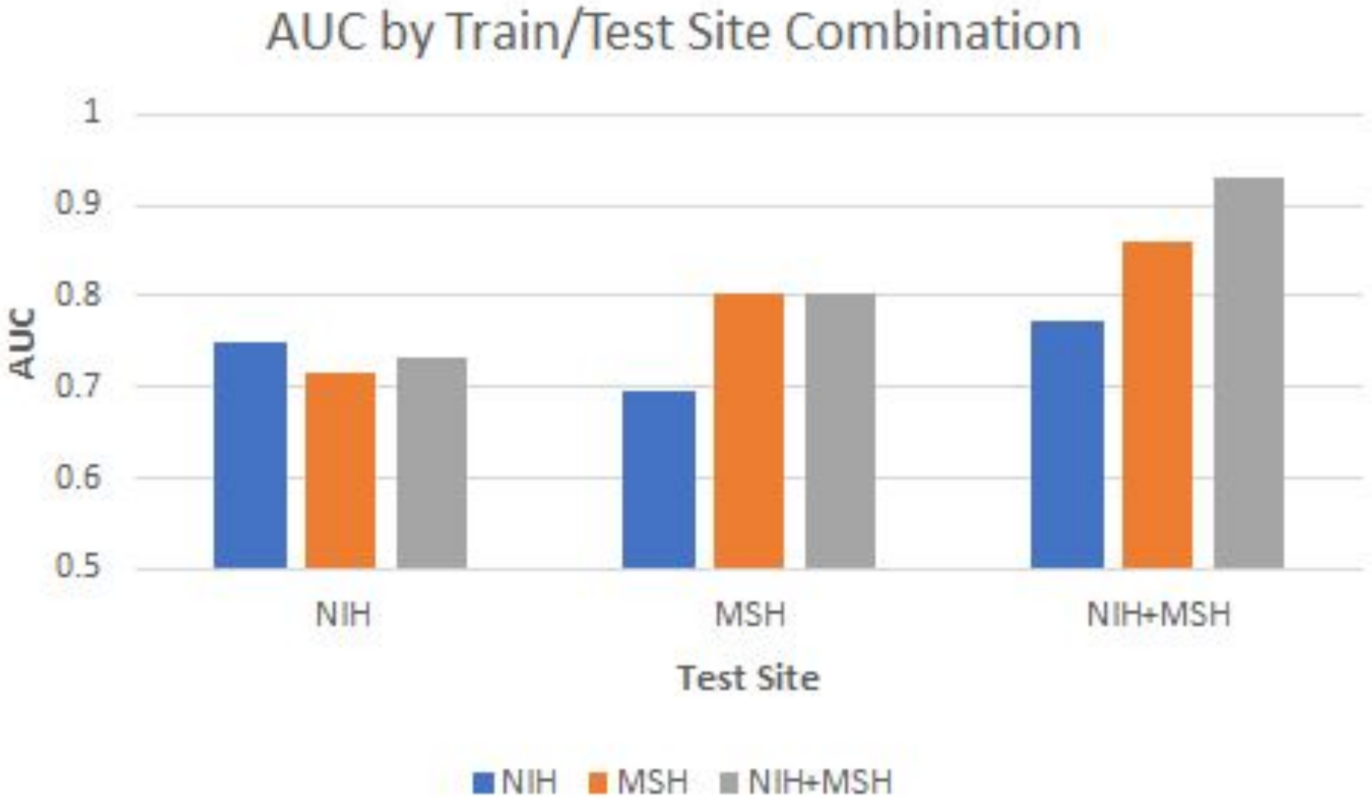
Confounders in Radiology: Zech et al. 2018

Table 3. Internal and external pneumonia screening performance for all train - tune and test hospital system combinations.

Train - Tune Site	Comparison Type*	Test Site (Images)	AUC (95% C.I.)	Acc.	Sens.	Spec.	PPV	NPV
NIH	Internal	NIH (N=22,062)	0.750 (0.721-0.778)	0.255	0.951	0.247	0.015	0.998
	External	MSH (N=8,388)	0.695 (0.683-0.706)	0.476	0.950	0.212	0.401	0.884
	External	IU (N=3,807)	0.725 (0.644-0.807)	0.190	0.974	0.182	0.012	0.999
	Superset	MSH + NIH (N=30,450)	0.773 (0.766-0.780)	0.462	0.950	0.403	0.160	0.985
	Superset	MSH + NIH + IU (N=34,257)	0.787 (0.780-0.793)	0.470	0.950	0.418	0.148	0.987
MSH	Internal	MSH (N=8,388)	0.802 (0.793-0.812)	0.617	0.950	0.432	0.482	0.940
	External	NIH (N=22,062)	0.717 (0.687-0.746)	0.184	0.951	0.175	0.014	0.997
	External	IU (N=3,807)	0.756 (0.674-0.838)	0.099	0.974	0.090	0.011	0.997
	Superset	MSH + NIH (N=30,450)	0.862 (0.856-0.868)	0.562	0.950	0.516	0.190	0.989
	Superset	MSH + NIH + IU (N=34,257)	0.871 (0.865-0.877)	0.577	0.950	0.537	0.180	0.990
MSH + NIH	Internal	MSH + NIH (N=30,450)	0.931 (0.927-0.936)	0.732	0.950	0.706	0.279	0.992
	Subset	NIH (N=22,062)	0.733 (0.703-0.762)	0.243	0.951	0.234	0.015	0.997
	Subset	MSH (N=8,388)	0.805 (0.796-0.814)	0.630	0.950	0.451	0.491	0.942
	External	IU (N=3,807)	0.815 (0.745-0.885)	0.238	0.974	0.230	0.013	0.999
	Superset	MSH + NIH + IU (N=34,257)	0.934 (0.929-0.938)	0.732	0.950	0.709	0.258	0.993

*Superset= a test dataset containing data from the same distribution (hospital system) as the training data as well as external data. Subset = a test dataset containing data from fewer distributions (hospital systems) then the training data.

Confounders in Radiology: Zech et al. 2018

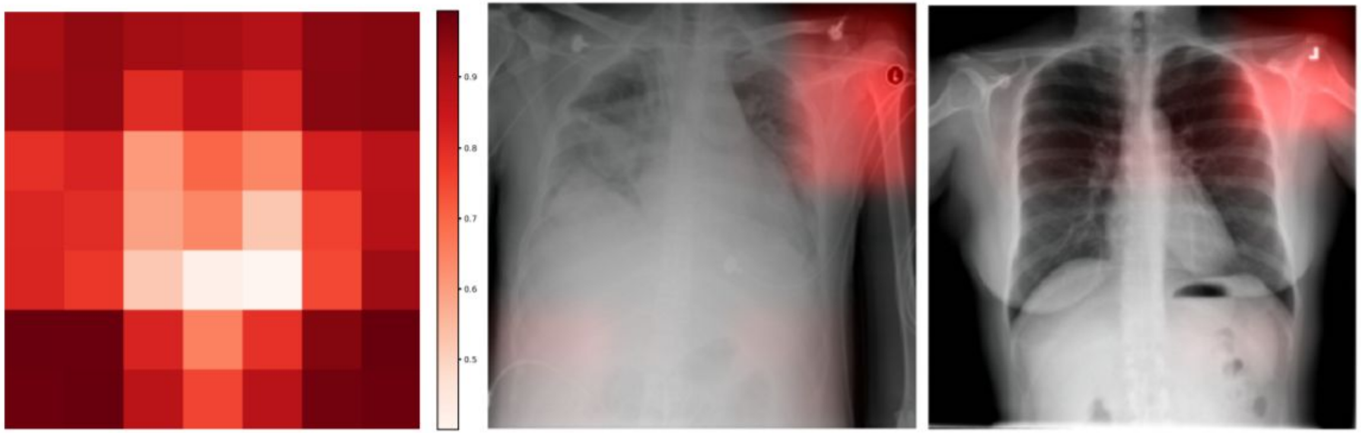


Why better performance on joint NIH+Mount Sinai dataset?

Confounders in Radiology: Zech et al. 2018

Why better performance on joint NIH+Mount Sinai dataset?

It’s learning to detect site: Mount Sinai has much higher pneumonia rate



(a) (b) (c)

Confounders in Radiology: Zech et al. 2018

Confounders in Radiology: Zech et al. 2018

- CNN learned something very useful in making predictions, but not clinically helpful.

Confounders in Radiology: Zech et al. 2018

- CNN learned something very useful in making predictions, but not clinically helpful.
- CNNs are hard to interpret: >6 million parameters

CNNs can an detect hospital system:

can it detect department within hospital?

CNNs can an detect hospital system:

can it detect department within hospital?

Yes.

At Mount Sinai, CNNs could detect portable x-ray scanner department (inpatient vs. ED) with near-perfect accuracy

We don’t have metadata for NIH, but...

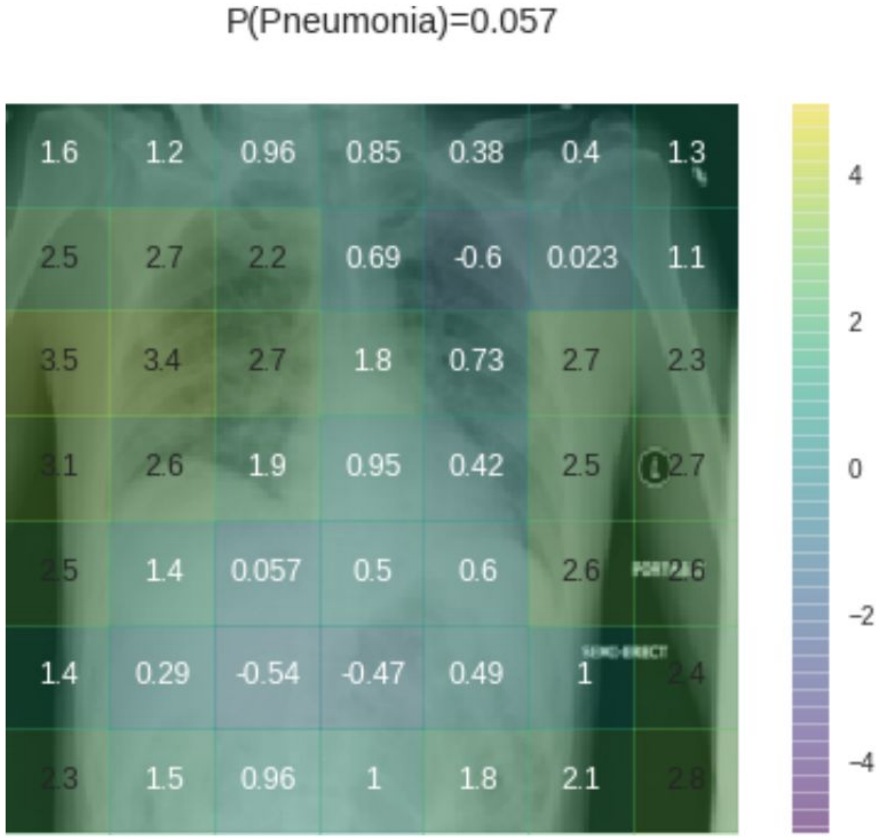
John Zech Follow
Preliminary medicine intern @CPMCinSF, future radiology resident @ColumbiaRadRes, passionate about machine learning. @johnrzech
Jul 8 · 9 min read

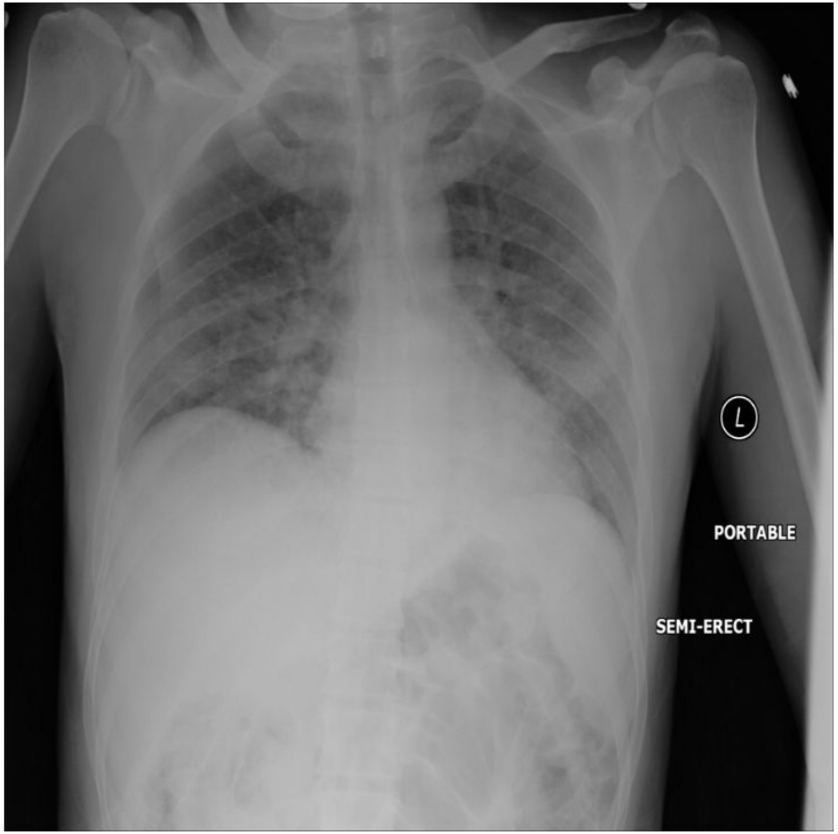
What are radiological deep learning models actually learning?

In radiology, we'd like deep learning models to identify patterns in imaging that suggest disease. For example, to detect pneumonia (lung infection), we'd like them to identify patterns in the lung that indicate the presence of an active infection. But do we know that is what they're actually doing?

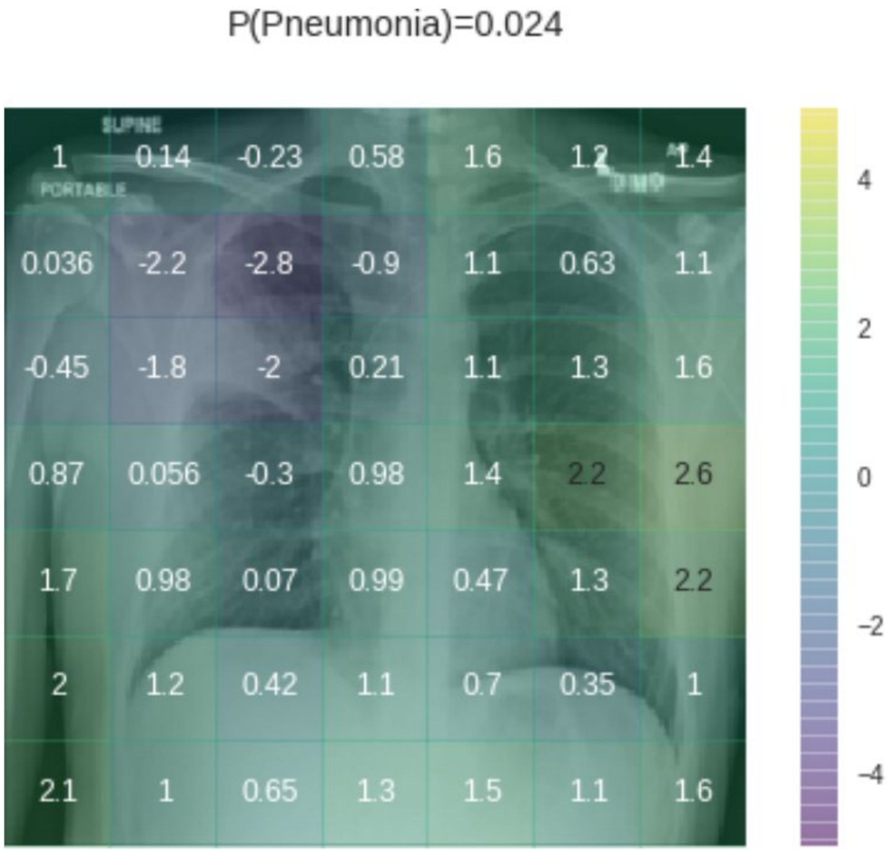
My collaborators and I recently released a preprint on arXiv examining how confounding variables may degrade the generalization performance of a CNN

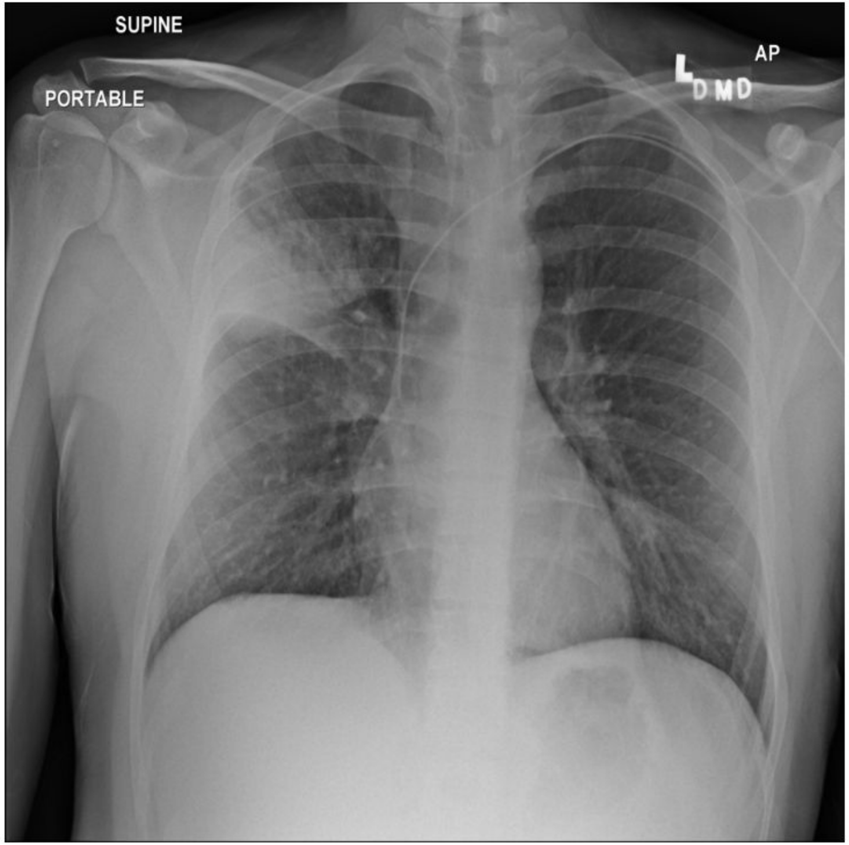
Confounders in Radiology: Zech (2018)



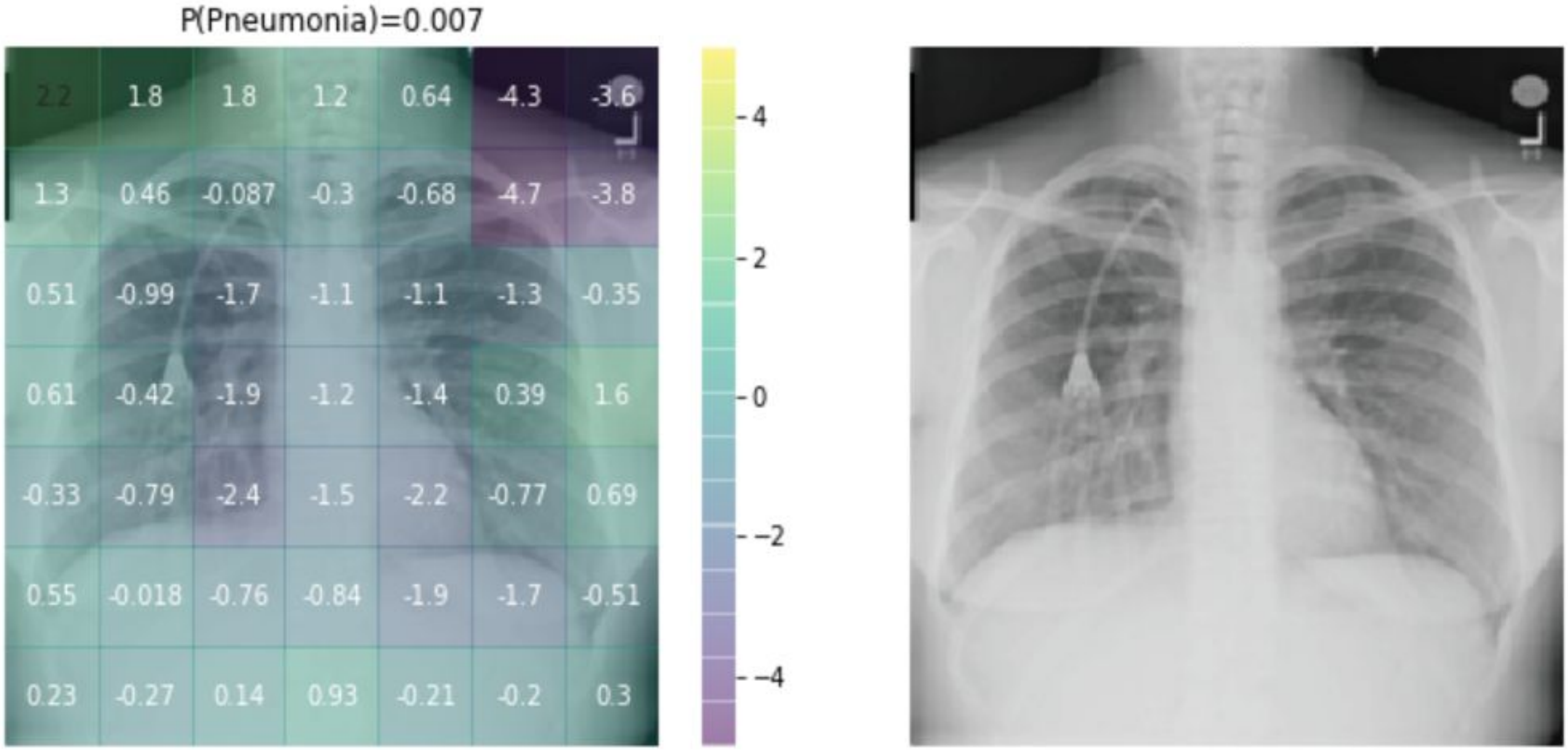


Confounders in Radiology: Zech (2018)





Confounders in Radiology: Zech (2018)



Confounders in Radiology: Zech et al. 2018

- **CNNs appear to exploit information beyond specific disease-related imaging findings on x-rays to calibrate their disease predictions.**

Confounders in Radiology: Zech et al. 2018

- **CNNs appear to exploit information beyond specific disease-related imaging findings on x-rays to calibrate their disease predictions**
- Scanner type (especially portable vs regular PA/lateral) is easily exploited

Confounders in Radiology: Zech et al. 2018

- **CNNs appear to exploit information beyond specific disease-related imaging findings on x-rays to calibrate their disease predictions**
- Scanner type (especially portable vs regular PA/lateral) is easily exploited
- Whole-image, low-res classification is especially vulnerable

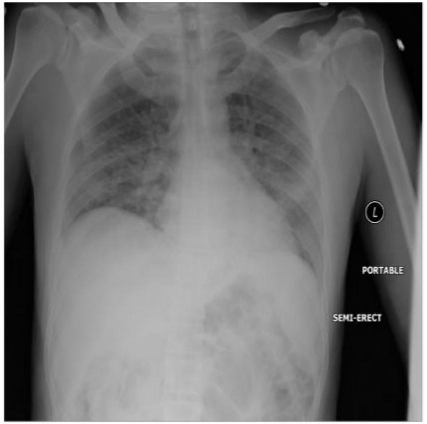
Confounders in Radiology: Zech et al. 2018

- **CNNs appear to exploit information beyond specific disease-related imaging findings on x-rays to calibrate their disease predictions**
- Scanner type (especially portable vs regular PA/lateral) is easily exploited
- Whole-image, low-res classification is especially vulnerable

Confounders in Radiology: Zech et al. 2018

- **CNNs appear to exploit information beyond specific disease-related imaging findings on x-rays to calibrate their disease predictions.**
- Scanner type (especially portable vs regular PA/lateral) is easily exploited.
- Whole-image, low-res classification is especially vulnerable

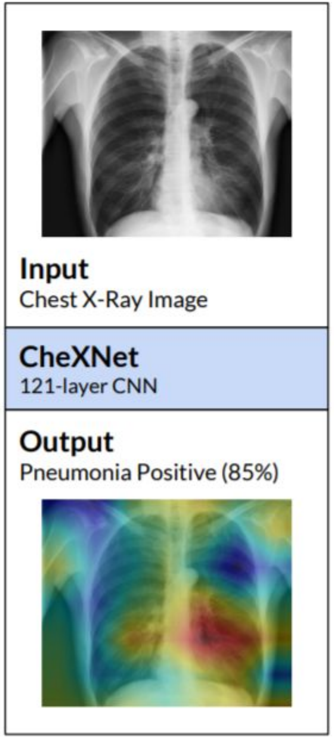
32 x 32



224 x 224

Rajpurkar et al. (2017)

- If the algorithm and the radiologists are given different tasks, is the comparison fair?
  - *Algorithm: use all information, including metadata implied by images, to optimize predictions*
  - *Radiologist: identify disease-specific findings*
- What does the ‘pneumonia’ label mean?
  - Remarkably low agreement among radiologists
  - Low accuracy of CNN
  - Imaging findings are REQUIRED for the diagnosis → raises questions given low inter-rater agreement



How do we move forward from weakly-supervised ImageNet-based transfer learning?

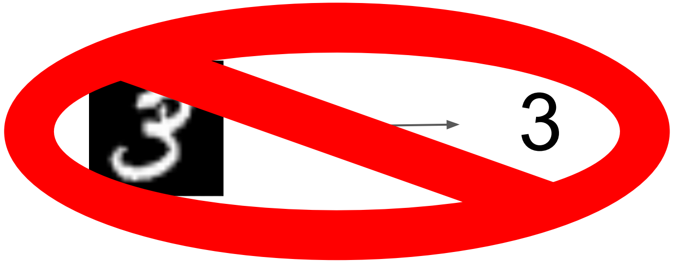
How do we move forward from weakly-supervised ImageNet-based transfer learning?

Domain adapted approaches that use segmentation

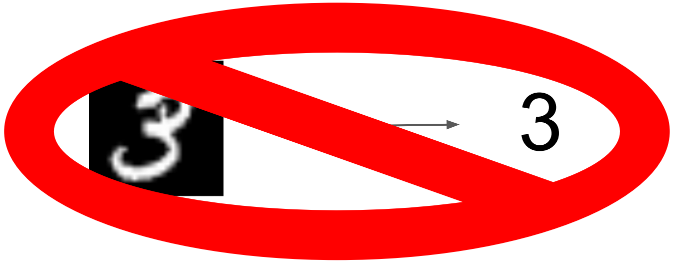
Domain-adapted CNN

3

Domain-adapted CNN



Domain-adapted CNN



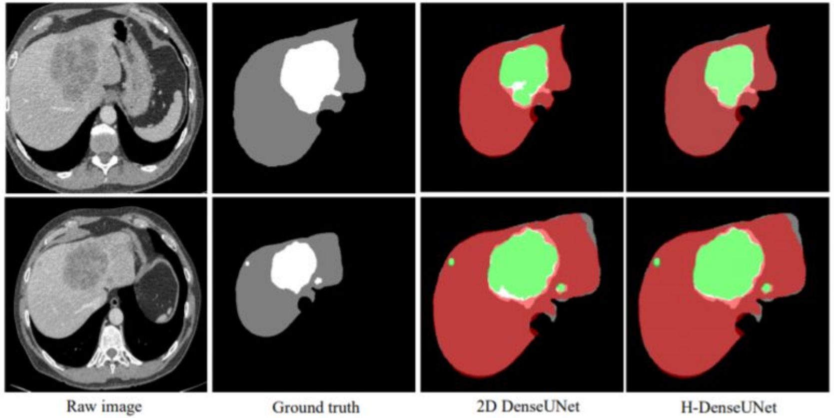
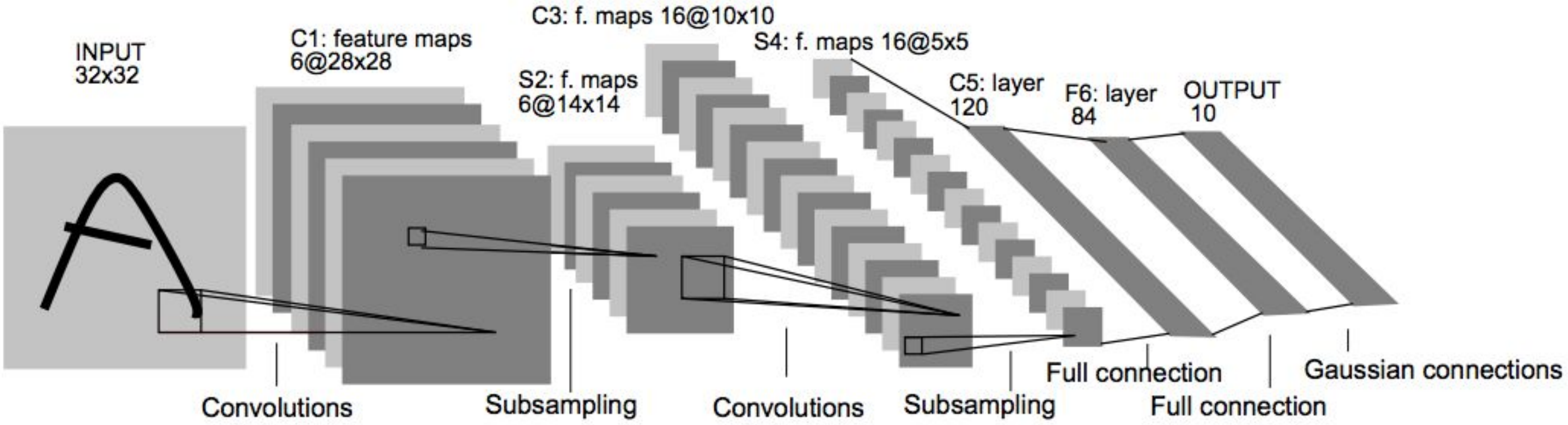


Figure 4: Examples of segmentation results by 2D DenseUNet and H-DenseUNet on the validation dataset. The *red* regions denote the segmented liver while the *green* ones denote the segmented lesions. The *gray* regions denote the true liver while the *white* ones denote the true lesions.

https://arxiv.org/pdf/1709.07330.pdf

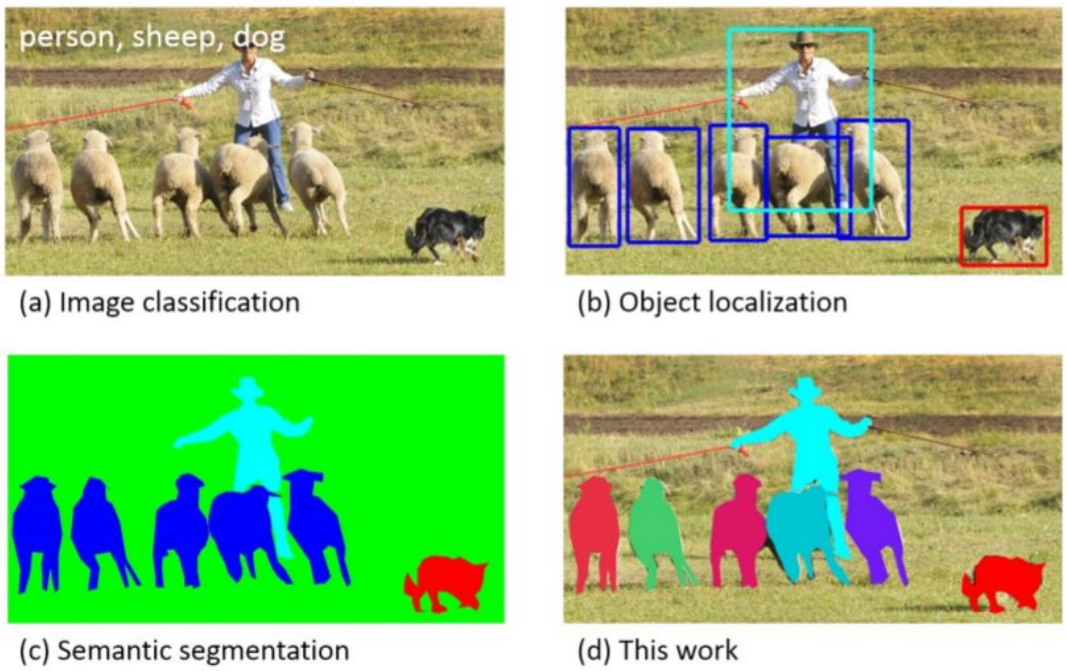
Classification CNNs

Maps image to a single classification



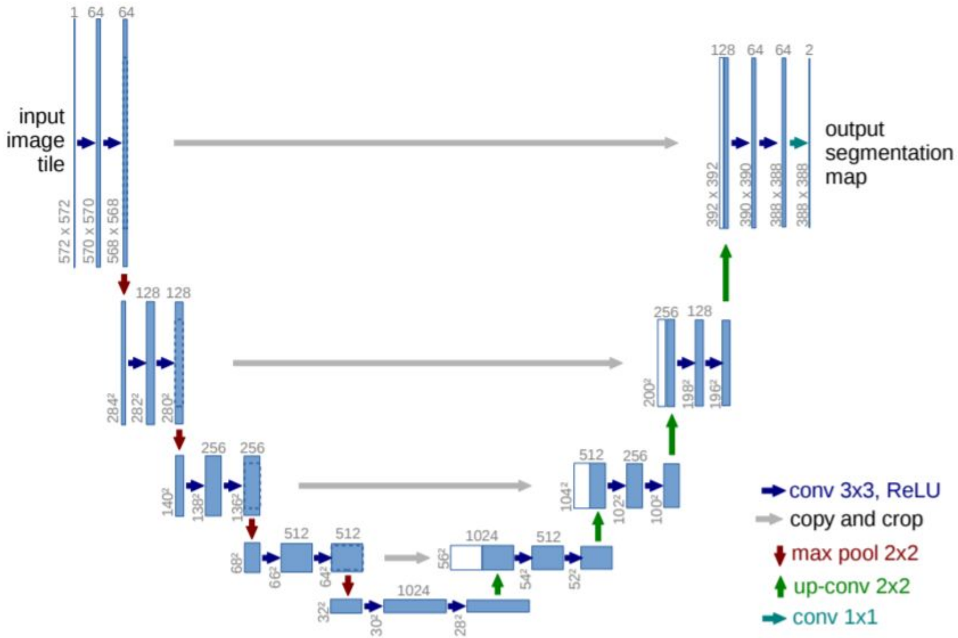
https://medium.com/@pechyonkin/key-deep-learning-architectures-lenet-5-6fc3c59e6f4

Segmentation CNNs



https://arxiv.org/pdf/1405.0312.pdf

Segmentation: U-Net



https://arxiv.org/pdf/1505.04597.pdf

Segmentation: U-Net

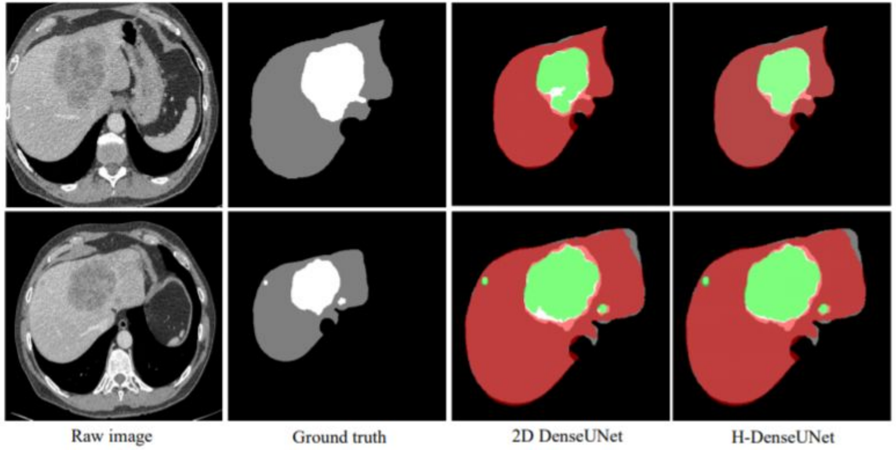
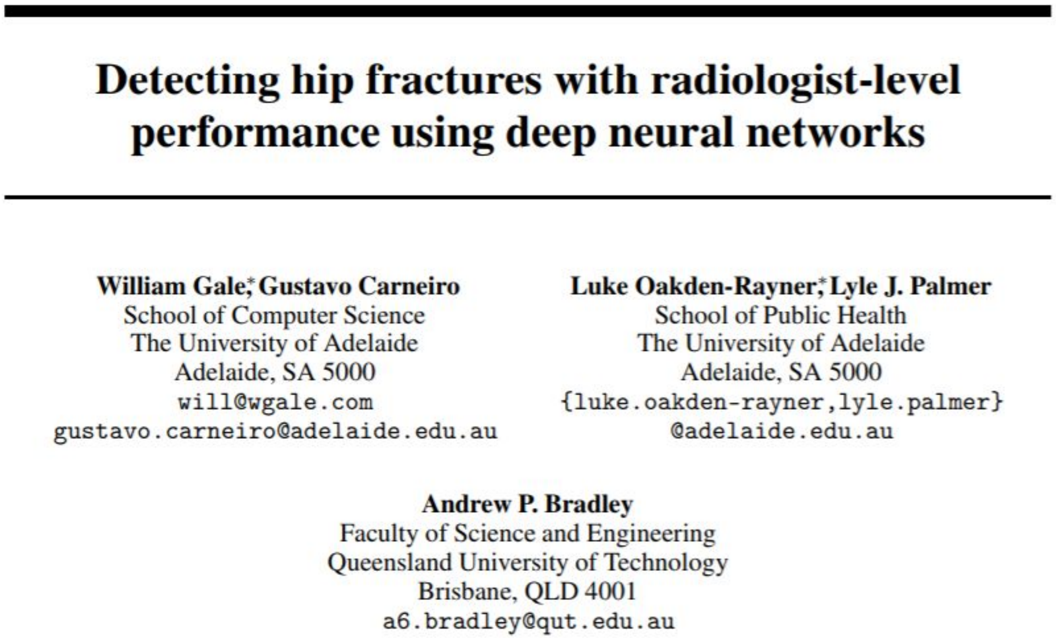


Figure 4: Examples of segmentation results by 2D DenseUNet and H-DenseUNet on the validation dataset. The *red* regions denote the segmented liver while the *green* ones denote the segmented lesions. The *gray* regions denote the true liver while the *white* ones denote the true lesions.

https://arxiv.org/pdf/1709.07330.pdf

Domain-adapted CNN: Gale et al. (2017)


Detecting hip fractures with radiologist-level performance using deep neural networks

William Gale*, Gustavo Carneiro
School of Computer Science
The University of Adelaide
Adelaide, SA 5000
will@wgale.com
gustavo.carneiro@adelaide.edu.au

Luke Oakden-Rayner*, Lyle J. Palmer
School of Public Health
The University of Adelaide
Adelaide, SA 5000
{luke.oakden-rayner,lyle.palmer}
@adelaide.edu.au

Andrew P. Bradley
Faculty of Science and Engineering
Queensland University of Technology
Brisbane, QLD 4001
a6.bradley@qut.edu.au

Domain-adapted CNN: Gale et al. (2017)

- Broad training dataset: 53,278 pelvis x-rays from Royal Adelaide Hospital
- Test set: only ED films

Domain-adapted CNN: Gale et al. (2017)

Four CNNs Used:

1. Filter for frontal x-rays
2. Locate head of femur: 1024 x 1024 pixels
3. Exclude films with metal implants
4. Customized DenseNet

Domain-adapted CNN: Gale et al. (2017)

Customized DenseNet

- 1024 x 1024 receptive field
- 1,434,176 parameters
- two loss functions
    - fracture/no fracture
    - Location: intra-capsular, extra-capsular, and no fracture

Domain-adapted CNN: Gale et al. (2017)

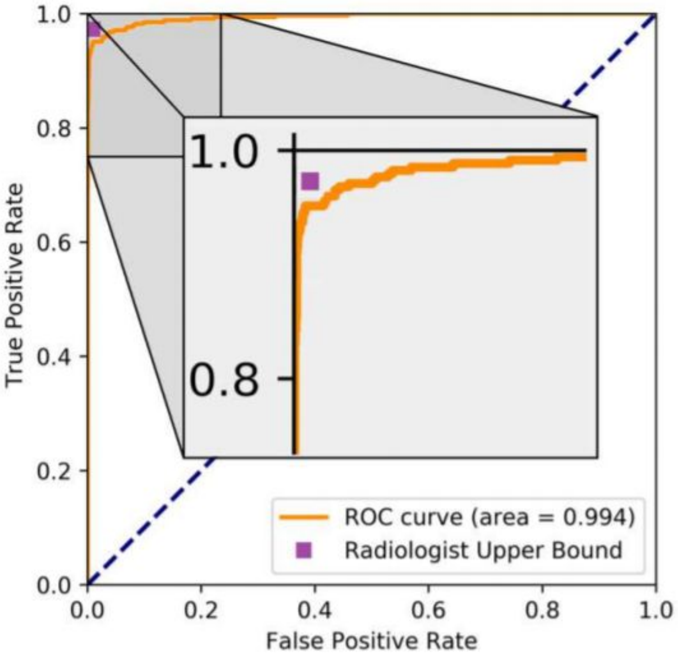


Figure 1: ROC curve showing the performance of the model with AUC 0.994, with a point reflecting the optimistic upper bound of human performance.

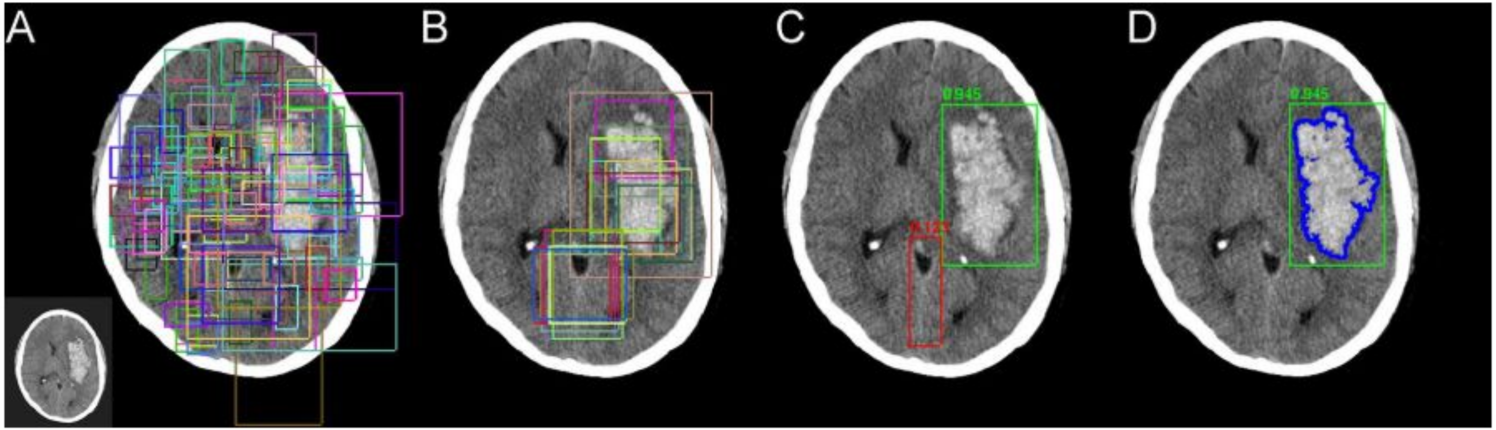
Domain-adapted CNN: Gale et al. (2017)

- Careful data cleaning to avoid confounding variables
  - Normalization
  - No metal
- Chosen test set reflecting real clinical use scenario: ED
- **Followed a radiologist’s process**
  - **zooming in on femur**
  - **maintain high resolution**

Domain-adapted CNN: Chang et al. (2018)

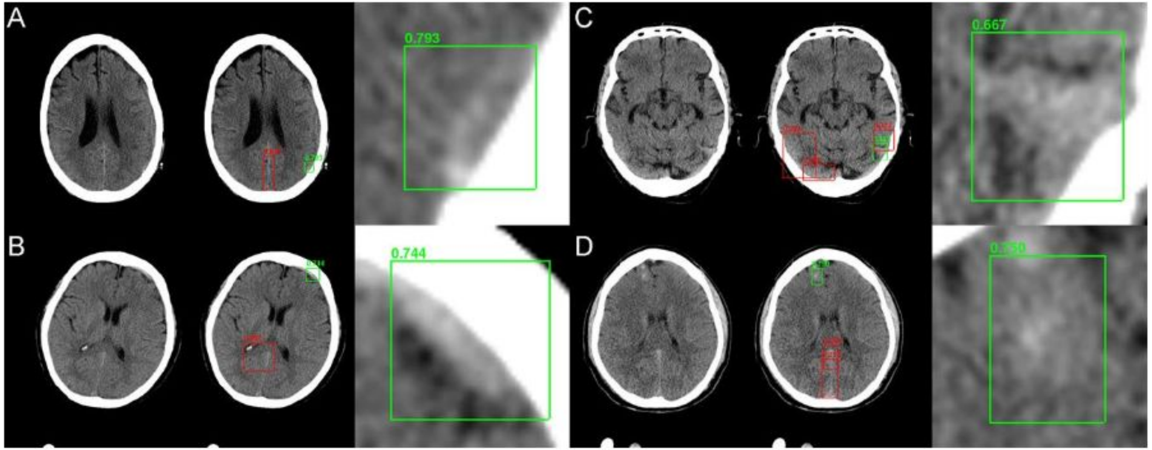
- Identify hemorrhage on 10,159 head CTs
- **Used segmentation-based approach**
- Results in challenging ED environment in true forward out of sample testing
  - 0.989 AUC
  - 97.2% accuracy
  - 0.951 sensitivity
  - 0.973 specificity

Domain-adapted CNN: Chang et al. (2018)



Mask residual CNN architectures can provide a framework for parallel evaluation of region proposal (attention), object detection (classification), and instance segmentation. In this approach, (A) preconfigured bounding boxes at various shapes and resolutions are tested for the presence of a potential abnormality. (B) The highest ranking bounding boxes are identified and used to generate region proposals that focus algorithm attention. (C) Composite region proposals are pruned using nonmaximum suppression and used as input into a classifier to determine presence or absence of hemorrhage. (D) Segmentation masks are generated for positive cases of hemorrhage. All images courtesy of Dr. Peter Chang.

Domain-adapted CNN: Chang et al. (2018)



Network predictions by the algorithm include bounding box region proposals for potential areas of abnormality (to focus algorithm attention) and final network predictions -- including confidence of the result. Correctly identified areas of hemorrhage (green) include subtle abnormalities representing subarachnoid (A), subdural (B and C), and intraparenchymal (D) hemorrhage. Correctly identified areas of excluded hemorrhage often include common mimics for blood on noncontrast CT including thickening/high density along the falx (A, B, and D) and beam hardening along the peripheral brain convexity (D).

Stronger approach and results,

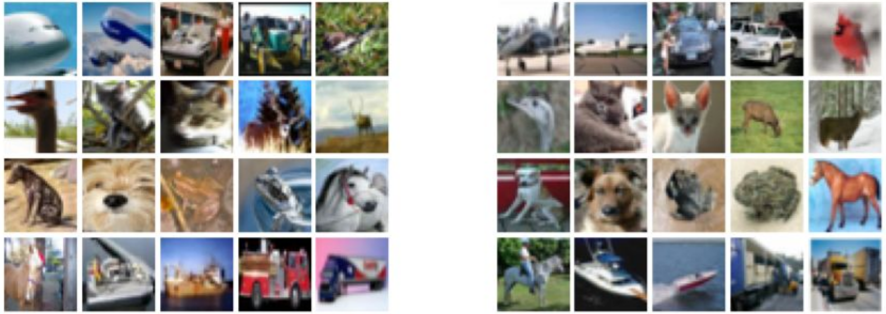
but needs generalization testing on new sites

Recht et al. (2018)

Do CIFAR-10 Classifiers Generalize to CIFAR-10?

Benjamin Recht, UC Berkeley
Rebecca Roelofs, UC Berkeley
Ludwig Schmidt, MIT
Vaishaal Shankar, UC Berkeley

June 4, 2018



(a) Test Set A (b) Test Set B

Figure 1: Class-balanced random draws from the new and original test sets.[1]

airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks

“Current accuracy numbers are brittle and susceptible to even minute natural variations in the data distribution.”

Recht et al. (2018)

How will we use deep learning in radiology?

How will we use deep learning in radiology?

- Can perform well at well-specified, clearly-designed imaging tasks: fracture, hemorrhage detection

How will we use deep learning in radiology?

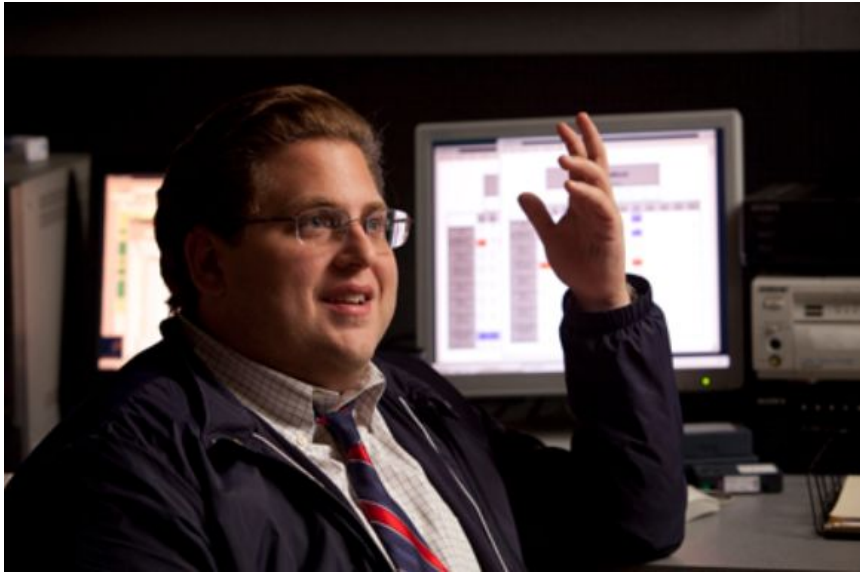
- Can perform well at well-specified, clearly-designed imaging tasks: fracture, hemorrhage detection
    - but must be carefully designed

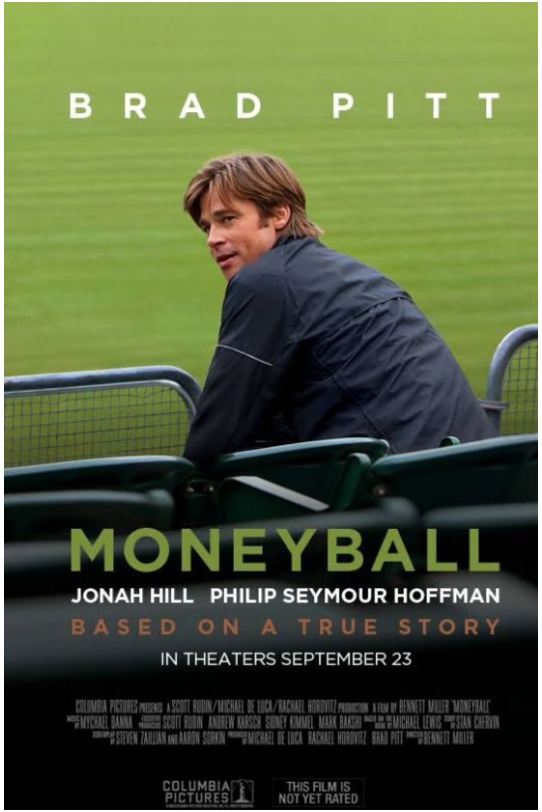
How will we use deep learning in radiology?

- Can perform well at well-specified, clearly-designed imaging tasks: fracture, hemorrhage detection
  - but must be carefully designed
- Could flag important information that affects interpretation, e.g., structured EHR data, text of physician notes

Are they truly ‘artificially intelligent’?

Or a (really intriguing) statistical model?





How will we combine this new information with our prior beliefs?

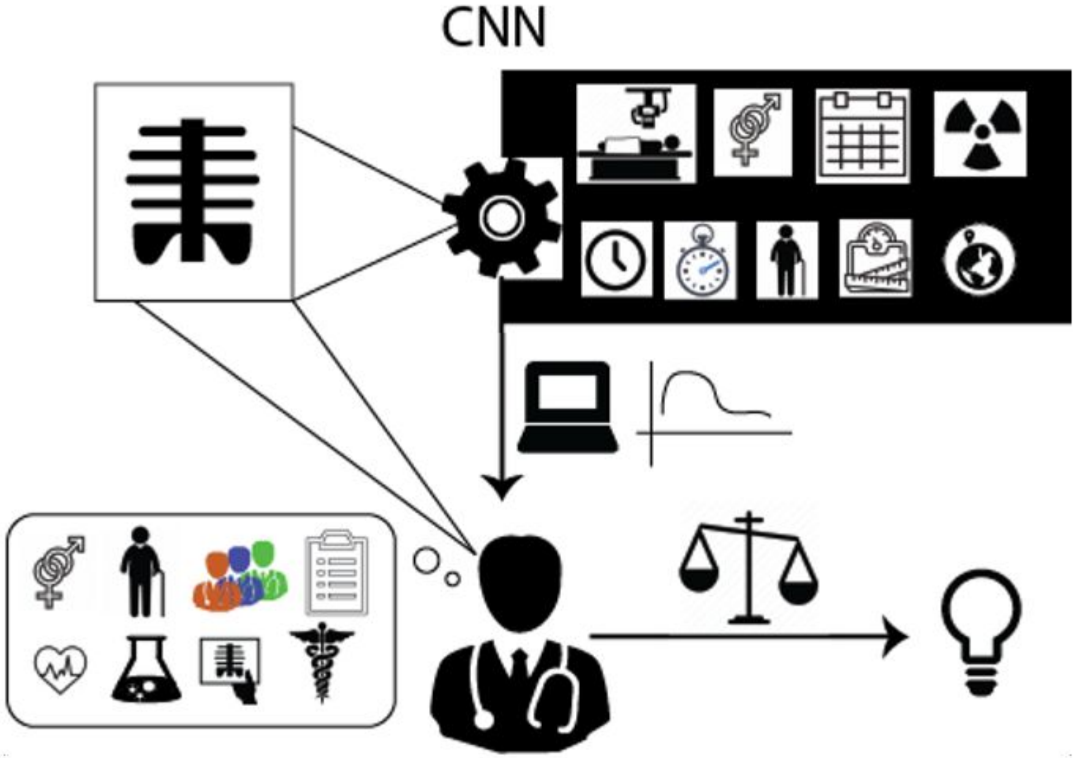
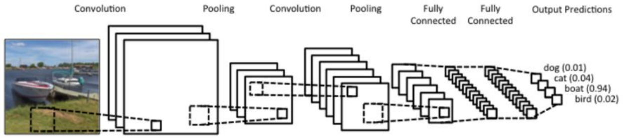


Figure courtesy Marcus Badgeley

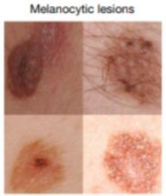
Takeaways

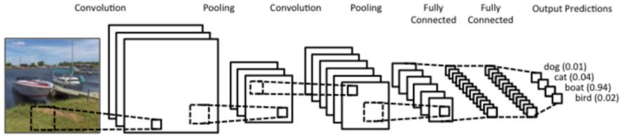
- What a convolutional neural network does



Takeaways

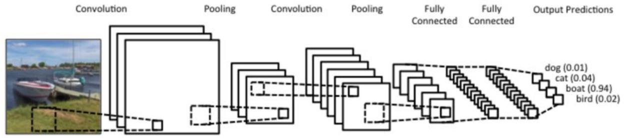
- What a convolutional neural network does
- Early promising results in dermatology



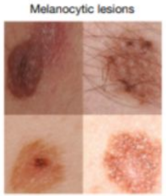


Takeaways

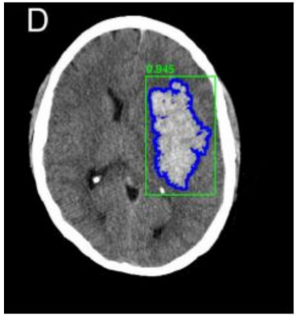
- What a convolutional neural network does

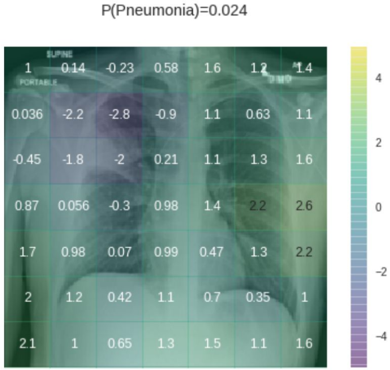


- Early promising results in dermatology



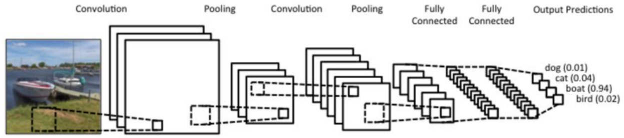
- Now used for weakly-supervised diagnosis in radiology, but **CNNs appear to exploit information beyond specific disease-related imaging findings on x-rays to calibrate their disease predictions**



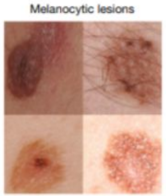


Takeaways

- What a convolutional neural network does

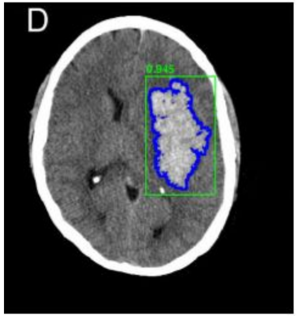


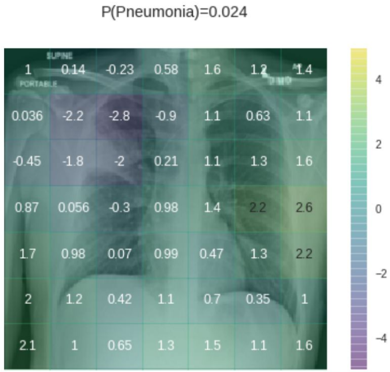
- Early promising results in dermatology



- Now used for weakly-supervised diagnosis in radiology, but **CNNs appear to exploit information beyond specific disease-related imaging findings on x-rays to calibrate their disease predictions**

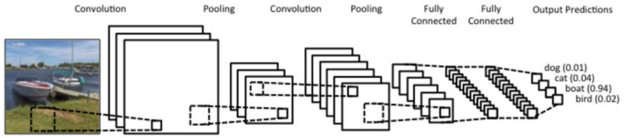
- Domain adapted approaches are promising, but generalization performance needs assessment



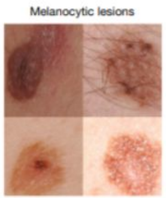


Takeaways

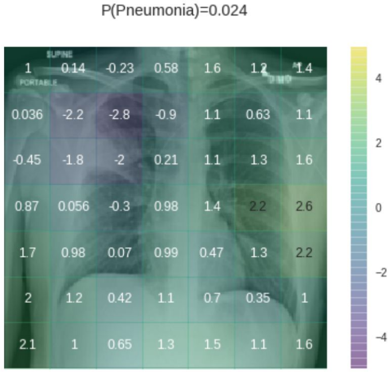
- What a convolutional neural network does



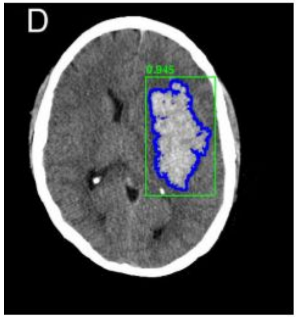
- Early promising results in dermatology



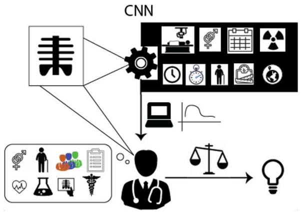
- Now used for weakly-supervised diagnosis in radiology, but **CNNs appear to exploit information beyond specific disease-related imaging findings on x-rays to calibrate their disease predictions**



- Domain adapted approaches are promising, but generalization performance needs assessment



- **And how to put it all together?**



Thank you!

Mount Sinai

UCSF

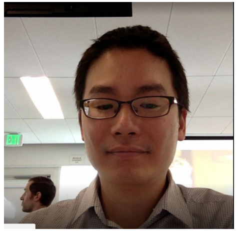
Sutter Health
CPMC











...and everyone else who contributed to these projects!