Lecture 13: Interpretability, Fairness, and Ethics

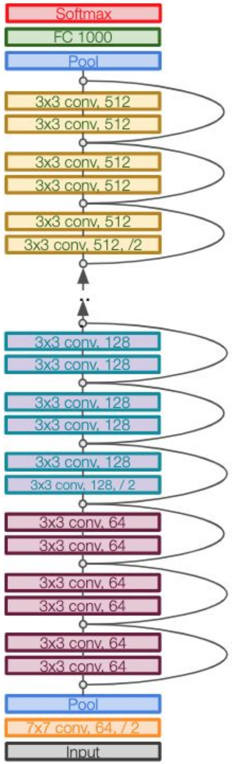
Announcements

- TA office hours next week will be used for project advising sessions instead
  - This is a required component of the project — sign up deadline is Fri 11/6
  - See Piazza post for details

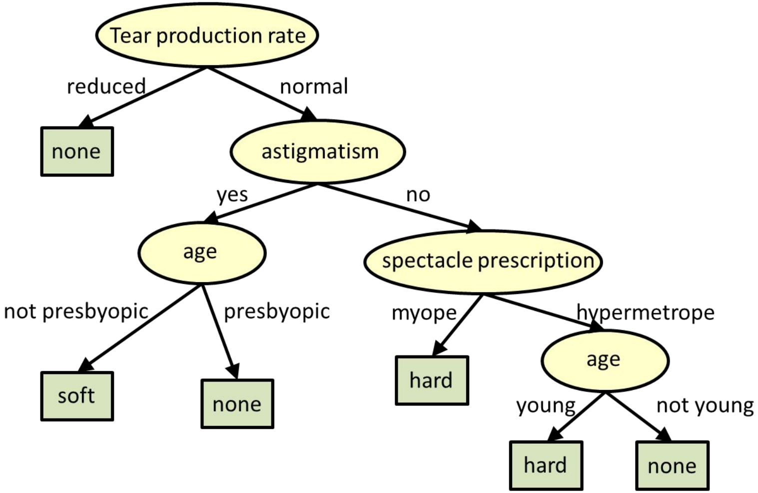
Many related concepts for today's lecture

- Interpretability
- Explainability
- Transparency
- Uncertainty
- Robustness
- Fairness
- Ethics
- Bias
- Etc...

Interpretability: a challenge in deep learning



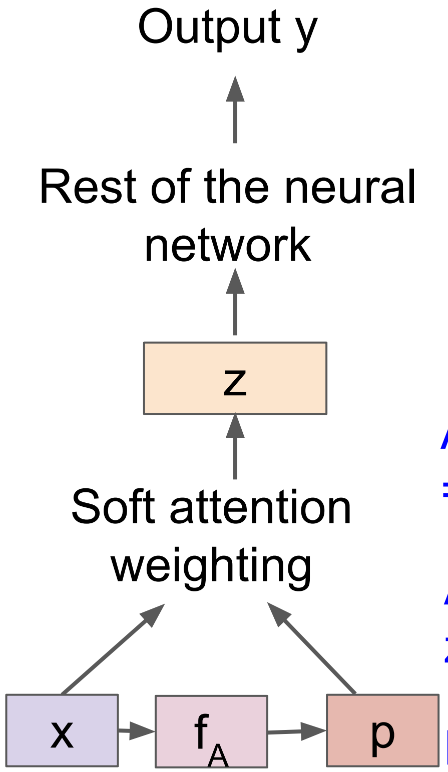
vs.



https://www.cs.cmu.edu/~bhiksha/courses/10-601/decisiontrees/DT.png

Soft attention: building interpretability into the model structure

- Weight input variables by an "attention weights" vector p
- Learn to dynamically produce p for any given input, by making it a function of the input x and a fully connected layer f_A (with learnable parameters A)
- By optimizing for prediction performance, network will learn to produce p that gives stronger weights to the most informative features in x!



$$p = \mathrm{softmax}(Ax)$$

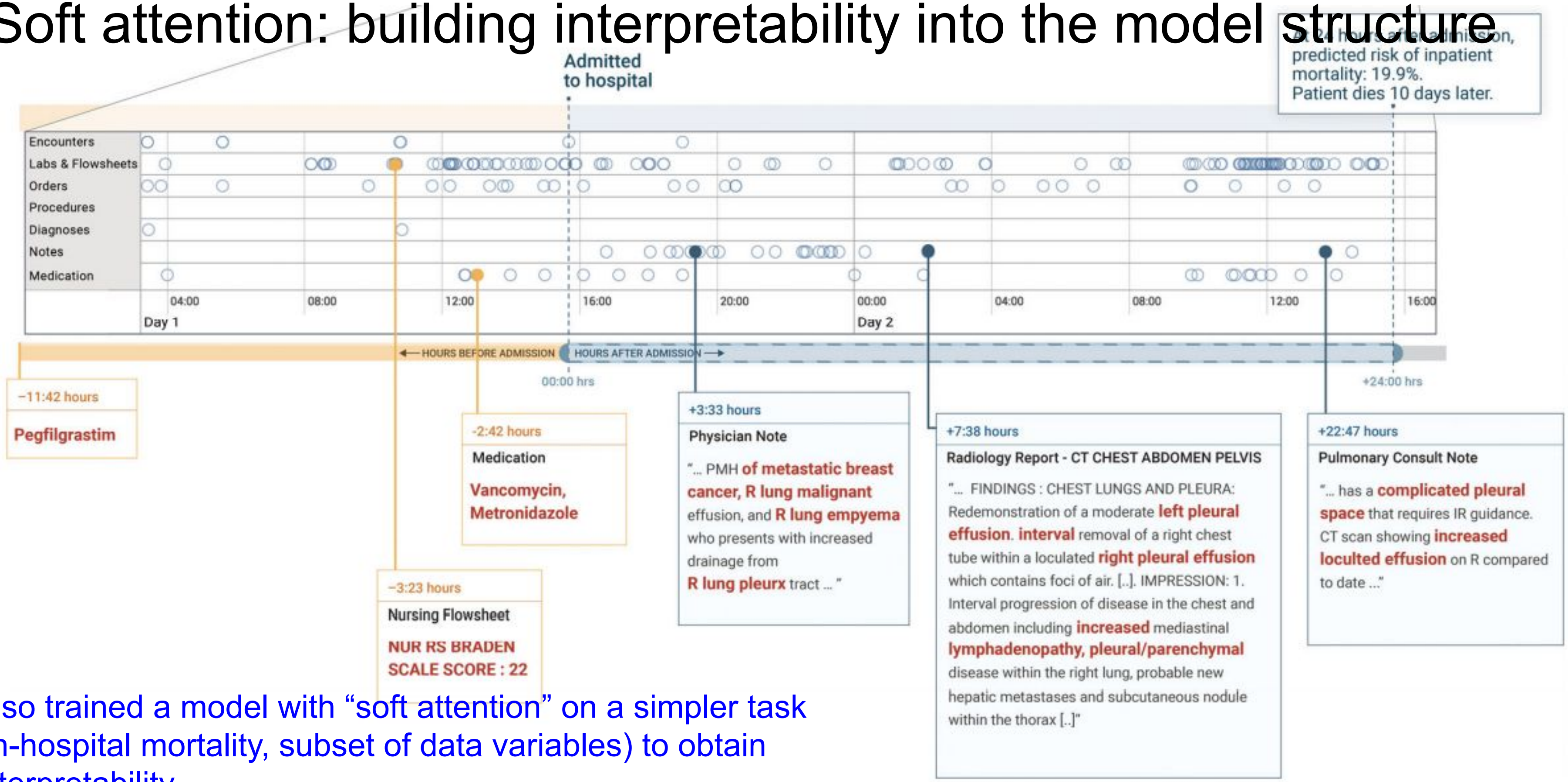
Input x = $[x_1, x_2, \ldots, x_D]$

Attention weights p $= [p_1, p_2, \ldots, p_D]$

Attention-weighted input z = $[z_1, z_2, \ldots, z_D]$

Learnable fully connected layer f_A with weights A

Soft attention: building interpretability into the model structure

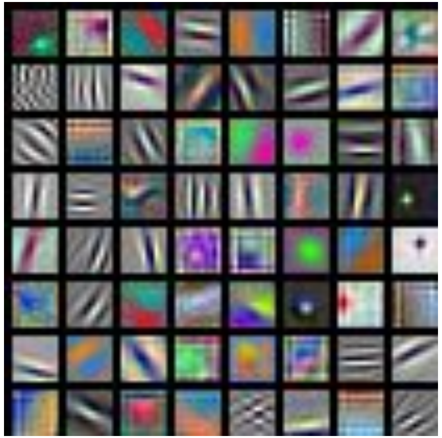


Also trained a model with "soft attention" on a simpler task (in-hospital mortality, subset of data variables) to obtain interpretability

Rajkomar et al. Scalable and accurate deep learning with electronic health records. Npj Digital Medicine, 2018.

How can we try to interpret a trained model?

First Layer: Visualize Filters

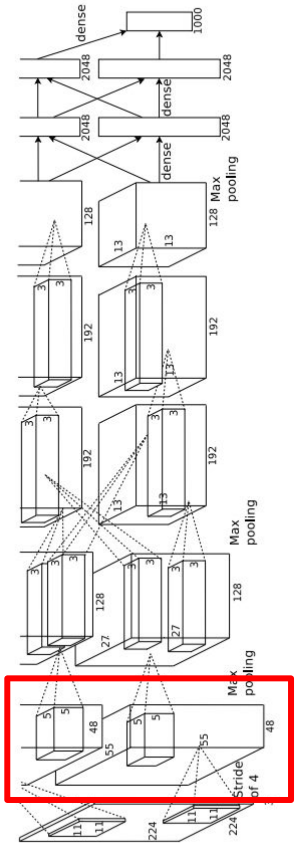


AlexNet:
64 x 3 x 11 x 11

ResNet-18:
64 x 3 x 7 x 7

ResNet-101:
64 x 3 x 7 x 7

DenseNet-121:
64 x 3 x 7 x 7



Krizhevsky, “One weird trick for parallelizing convolutional neural networks”, arXiv 2014
He et al, “Deep Residual Learning for Image Recognition”, CVPR 2016
Huang et al, “Densely Connected Convolutional Networks”, CVPR 2017

Slide credit: CS231n

Visualize the filters/kernels (raw weights)

We can visualize filters at higher layers, but not that interesting

(these are taken from ConvNetJS CIFAR-10 demo)

layer 1 weights

16 x 3 x 7 x 7

layer 2 weights

20 x 16 x 7 x 7

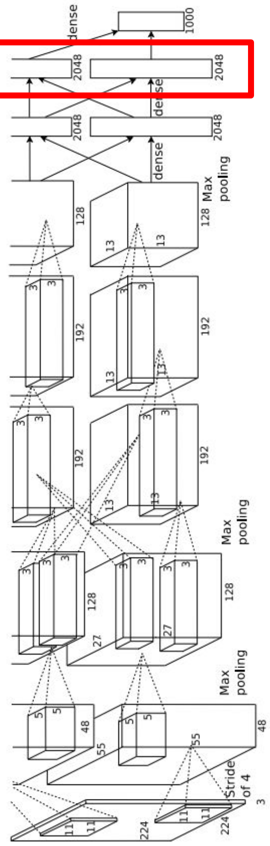
layer 3 weights

20 x 20 x 7 x 7

Slide credit: CS231n

Last Layer

FC7 layer



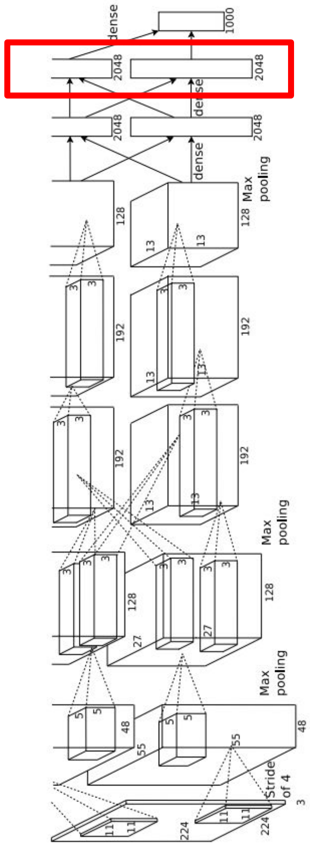
4096-dimensional feature vector for an image
(layer immediately before the classifier)

Run the network on many images, collect the feature vectors

Slide credit: CS231n

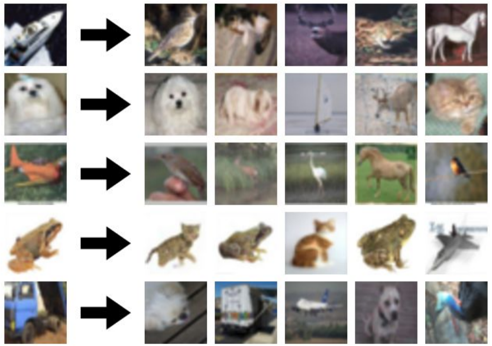
Last Layer: Nearest Neighbors

4096-dim vector



Test image   L2 Nearest neighbors in <u>feature</u> space

Recall: Nearest neighbors in <u>pixel</u> space



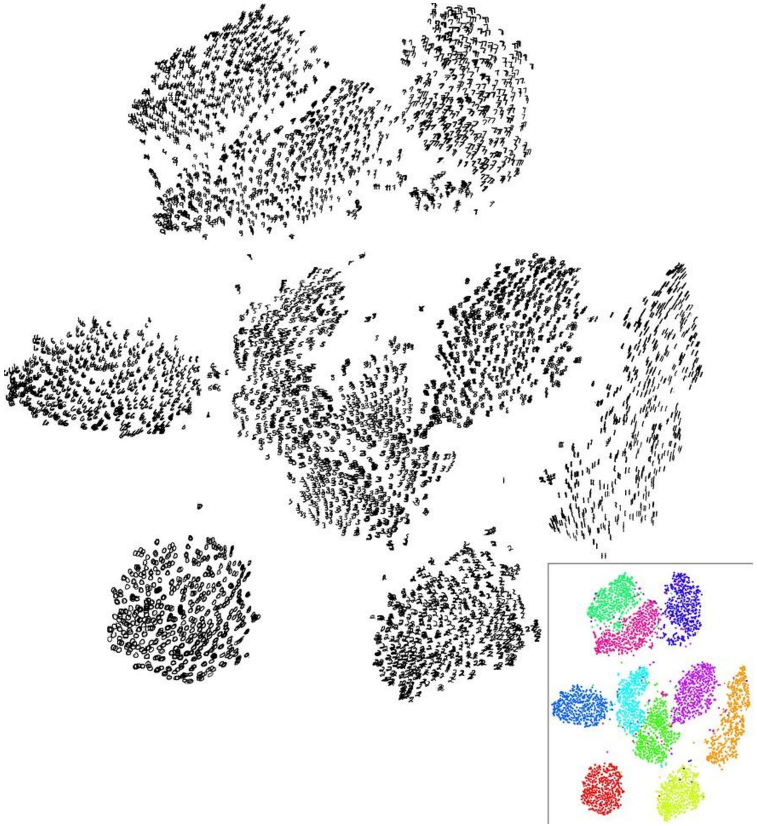
Krizhevsky et al, “ImageNet Classification with Deep Convolutional Neural Networks”, NIPS 2012.
Figures reproduced with permission.

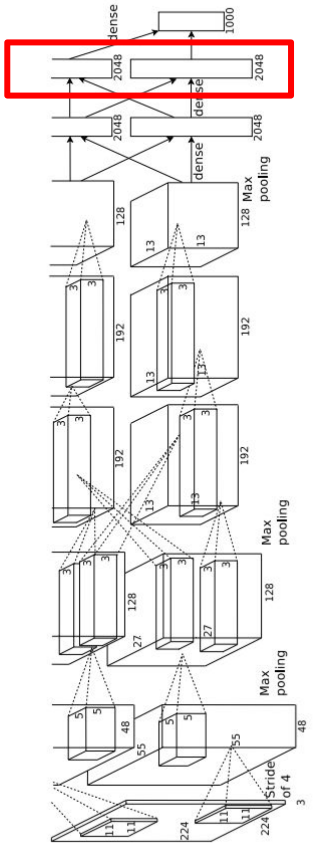
Slide credit: CS231n

Last Layer: Dimensionality Reduction

Visualize the “space” of FC7 feature vectors by reducing dimensionality of vectors from 4096 to 2 dimensions

Common approaches: **t-SNE and UMAP**

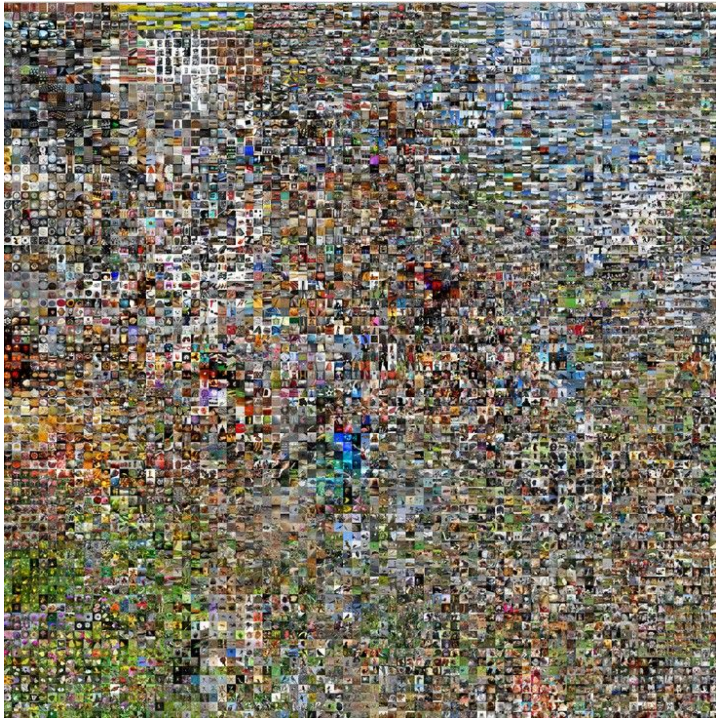


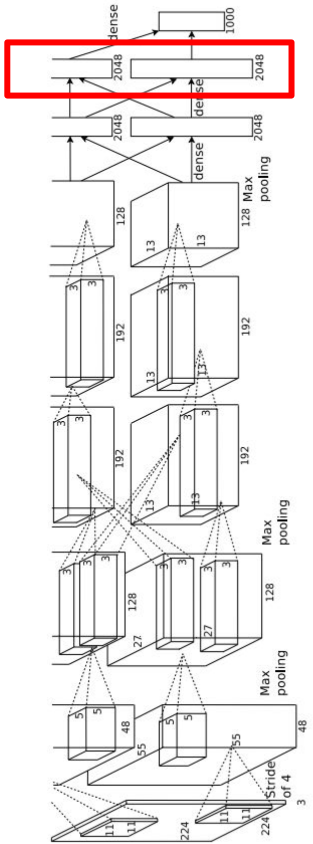


Van der Maaten and Hinton, “Visualizing Data using t-SNE”, JMLR 2008

Slide credit: CS231n

Last Layer: Dimensionality Reduction





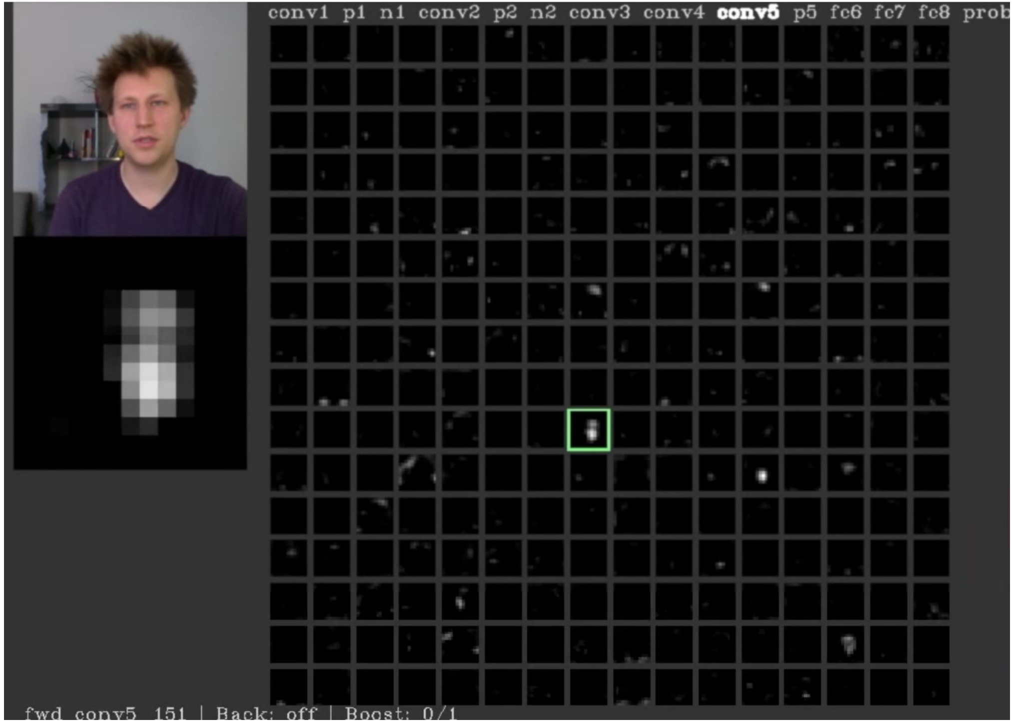
Van der Maaten and Hinton, “Visualizing Data using t-SNE”, JMLR 2008
Krizhevsky et al, “ImageNet Classification with Deep Convolutional Neural Networks”, NIPS 2012.
Figure reproduced with permission.

See high-resolution versions at
http://cs.stanford.edu/people/karpathy/cnnembed/

Slide credit: CS231n

Visualizing Activations

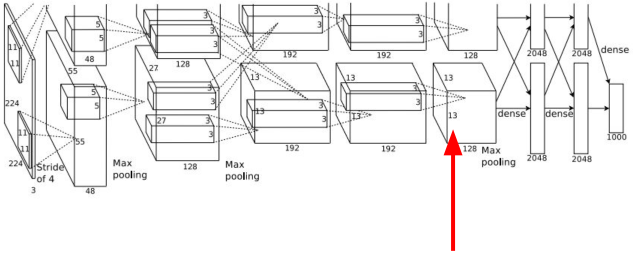
conv5 feature map is 128x13x13; visualize as 128 13x13 grayscale images

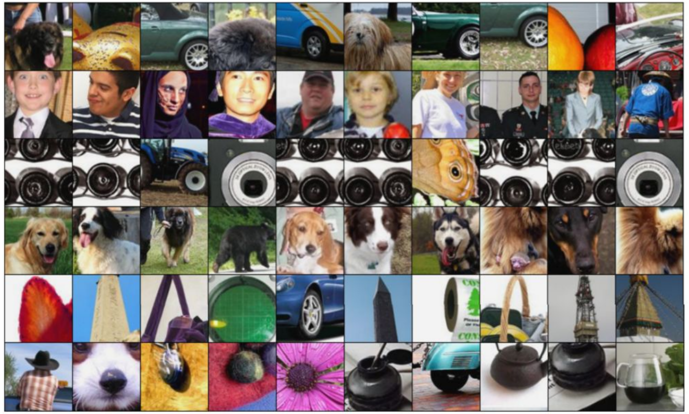


Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014.

Slide credit: CS231n

Maximally Activating Patches





Pick a layer and a channel; e.g. conv5 is 128 x 13 x 13, pick channel 17/128

Run many images through the network, record values of chosen channel

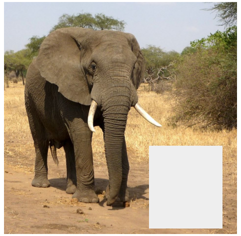
Visualize image patches that correspond to maximal activations

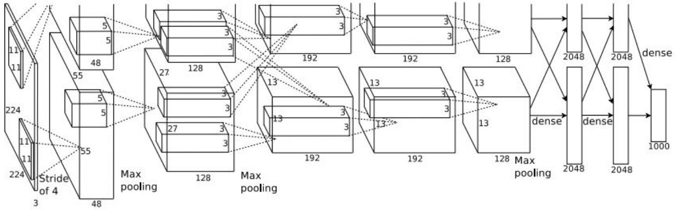
Springenberg et al, “Striving for Simplicity: The All Convolutional Net”, ICLR Workshop 2015

Slide credit: CS231n

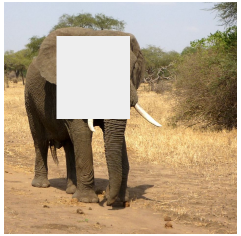
Which pixels matter: Saliency via Occlusion

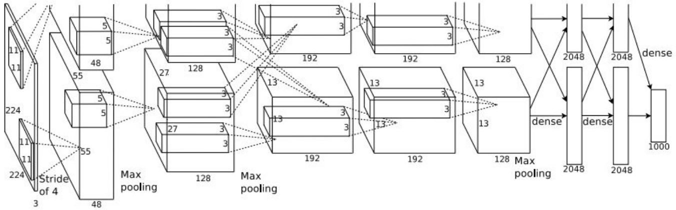
Mask part of the image before feeding to CNN, check how much predicted probabilities change





P(elephant) = 0.95





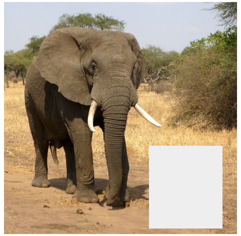
P(elephant) = 0.75

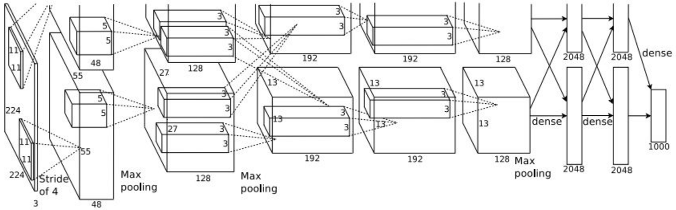
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

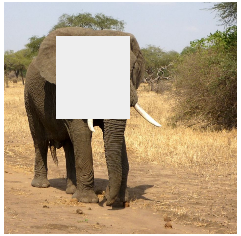
Slide credit: CS231n

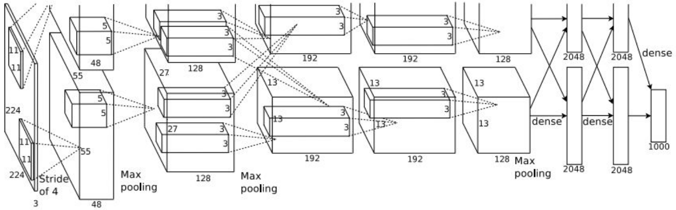
Which pixels matter: Saliency via Occlusion

Mask part of the image before feeding to CNN, check how much predicted probabilities change

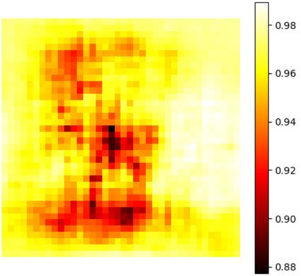




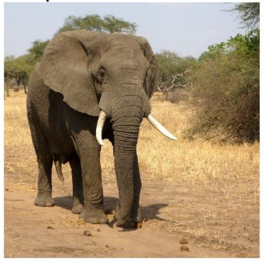


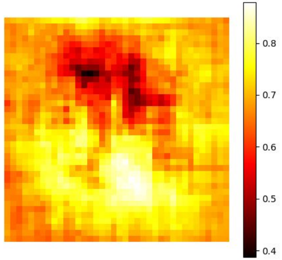


schooner

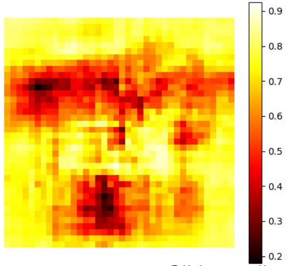


African elephant, Loxodonta africana





go-kart

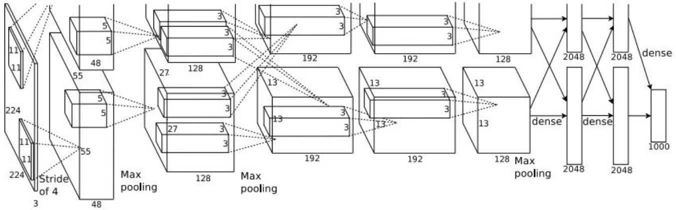


Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Slide credit: CS231n

Which pixels matter: Saliency via Backprop

Forward pass: Compute probabilities



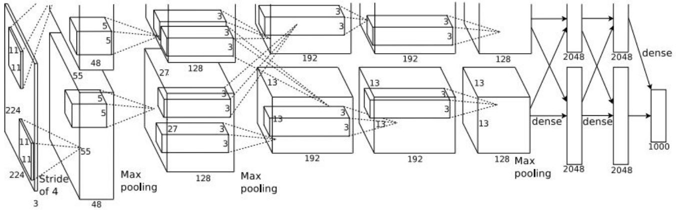
Dog

Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014.

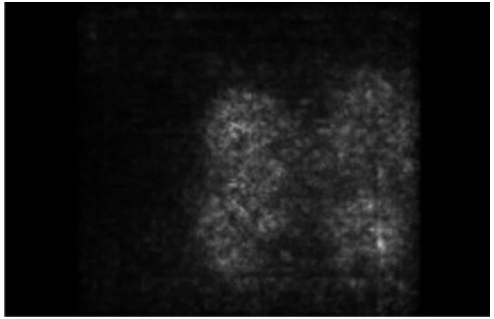
Slide credit: CS231n

Which pixels matter: Saliency via Backprop

Forward pass: Compute probabilities



Dog

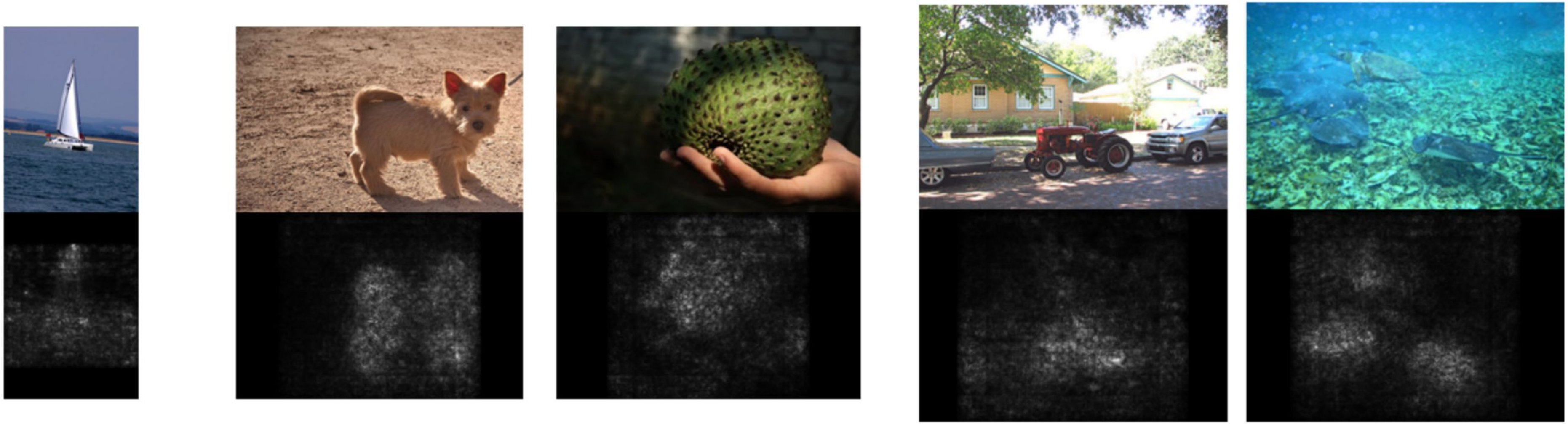


Compute gradient of (unnormalized) class score with respect to image pixels, take absolute value and max over RGB channels

Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014.

Slide credit: CS231n

Saliency Maps

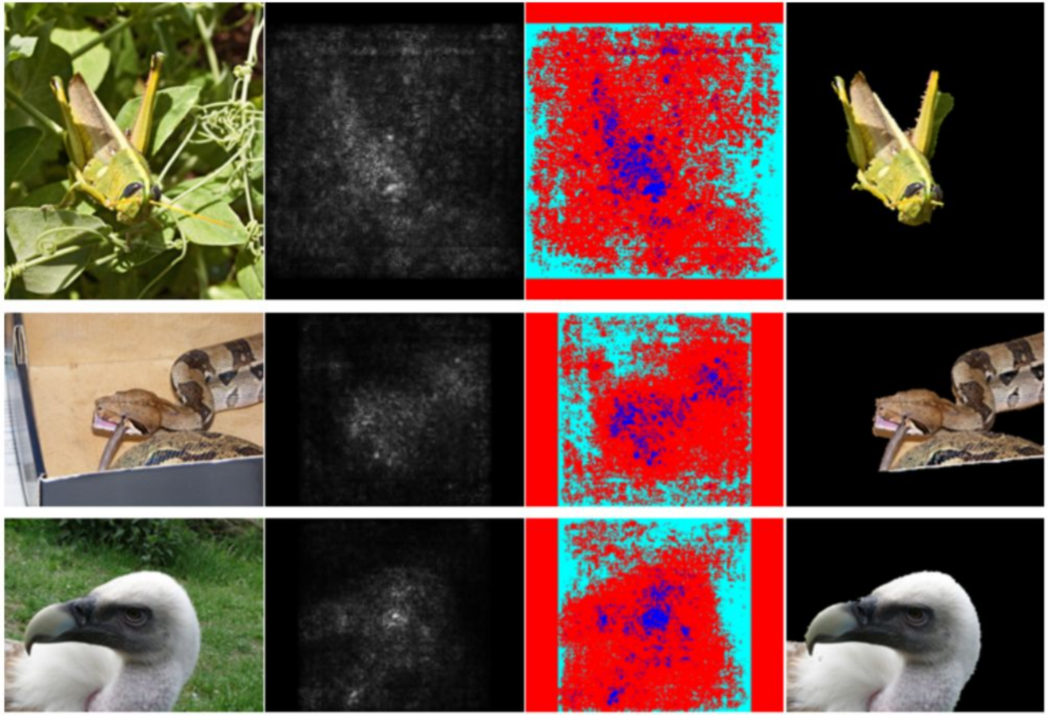


Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014.

Slide credit: CS231n

Saliency Maps: Segmentation without supervision

Use GrabCut on saliency map



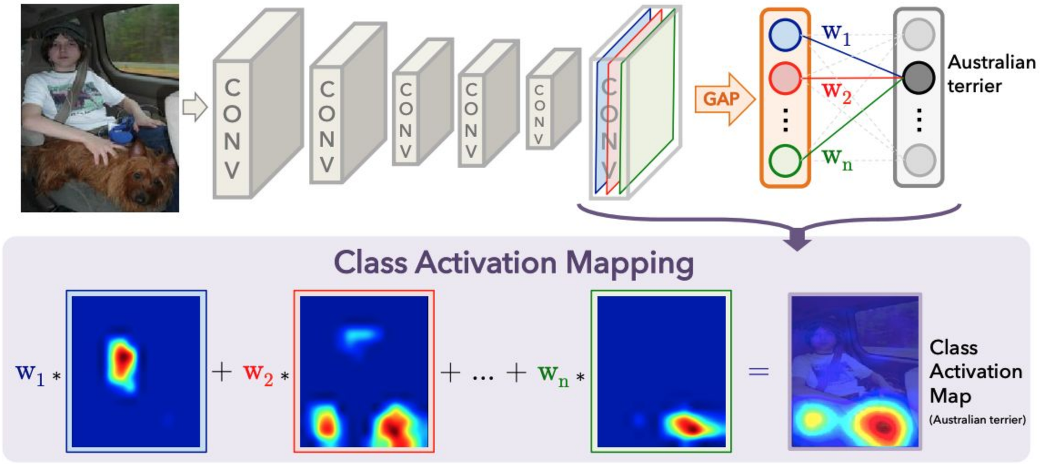
Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014.

Rother et al, “Grabcut: Interactive foreground extraction using iterated graph cuts”, ACM TOG 2004

Slide credit: CS231n

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.



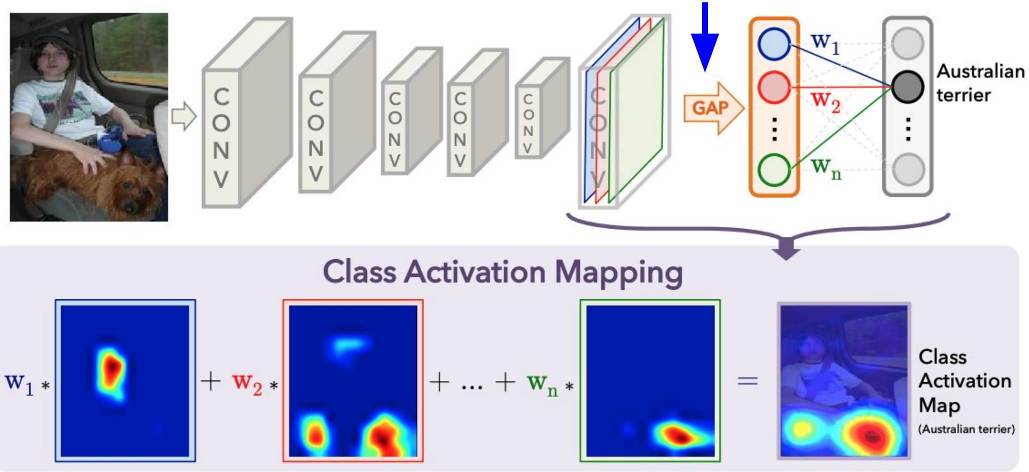
$$M_c(x, y) = \sum_k w_k^c f_k(x, y).$$

Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.

Relies on idea that global average pooling layers aggregate "signal" for particular patterns



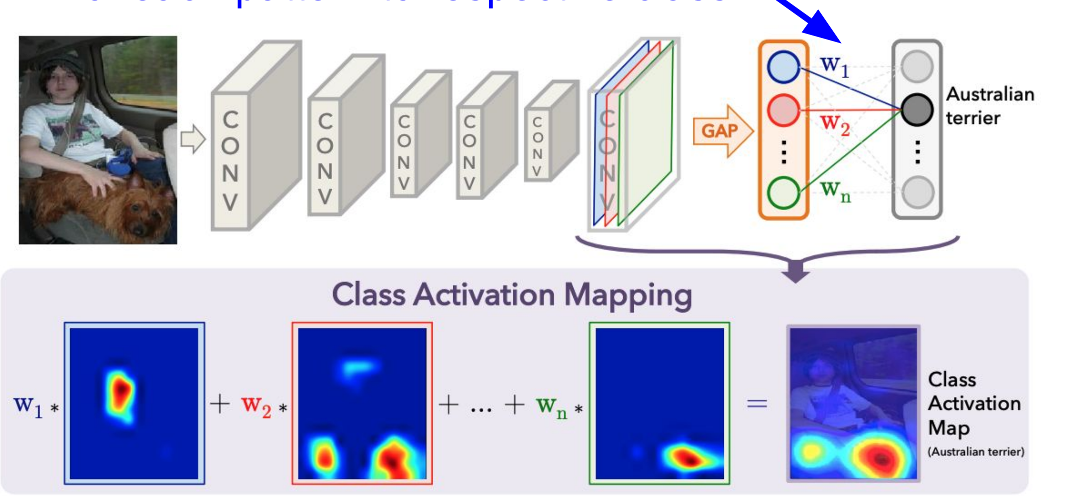
$$M_c(x, y) = \sum_k w_k^c f_k(x, y).$$

Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.

Weights of final classification layer gives importance of each pattern to respective class

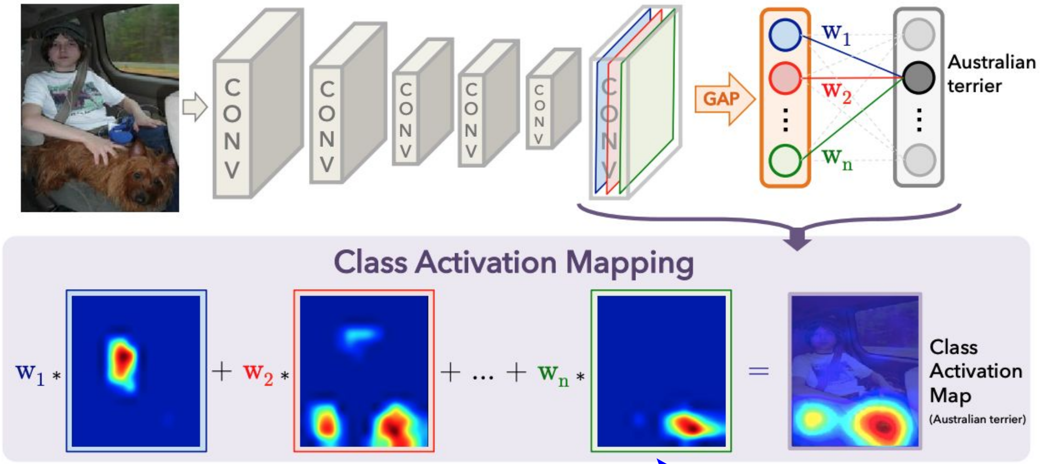


$$M_c(x,y) = \sum_k w_k^c f_k(x,y).$$

Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.



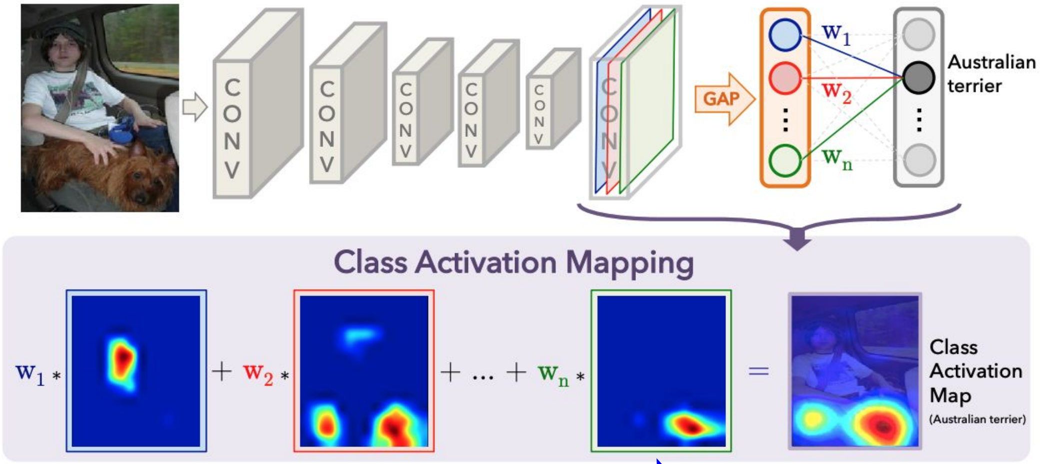
$$M_c(x, y) = \sum_k w_k^c f_k(x, y).$$

Heat maps before GAP show the spatially localized signal that will be aggregated

Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.



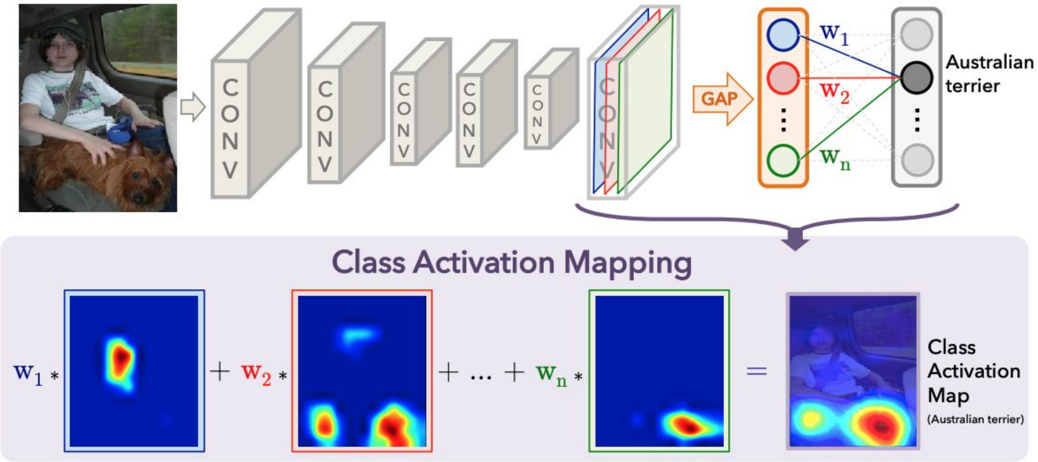
$$M_c(x, y) = \sum_k w_k^c f_k(x, y).$$

Add up all activation maps, weighted by importance to class (w_1,...,w_n) to get CAM

Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.

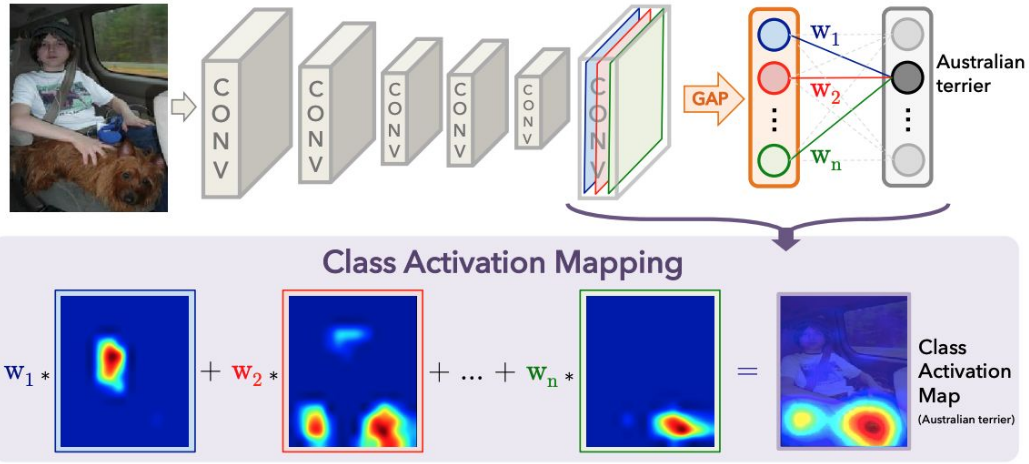


CAM → $$M_c(x, y) = \sum_k w_k^c f_k(x, y).$$

Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.



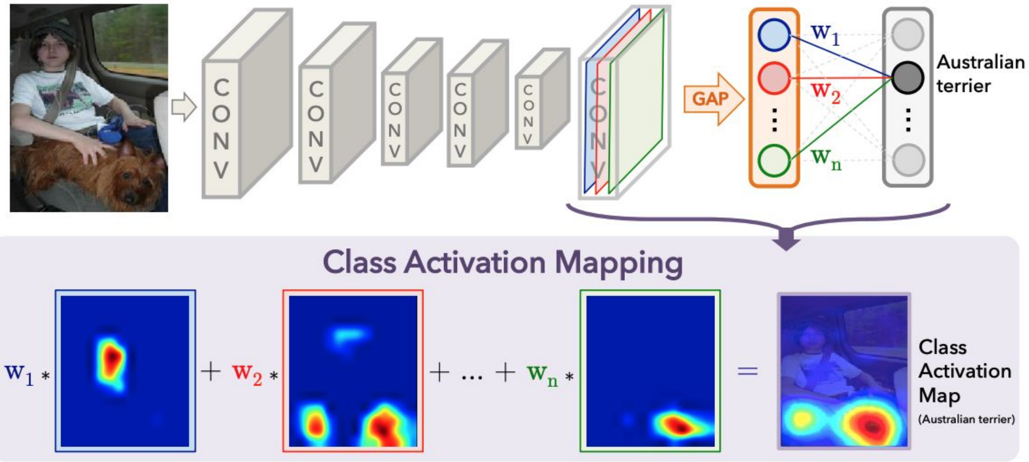
$$M_c(x,y) = \sum_k w_k^c f_k(x,y).$$

Activation map for kth filter in layer before GAP

Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.



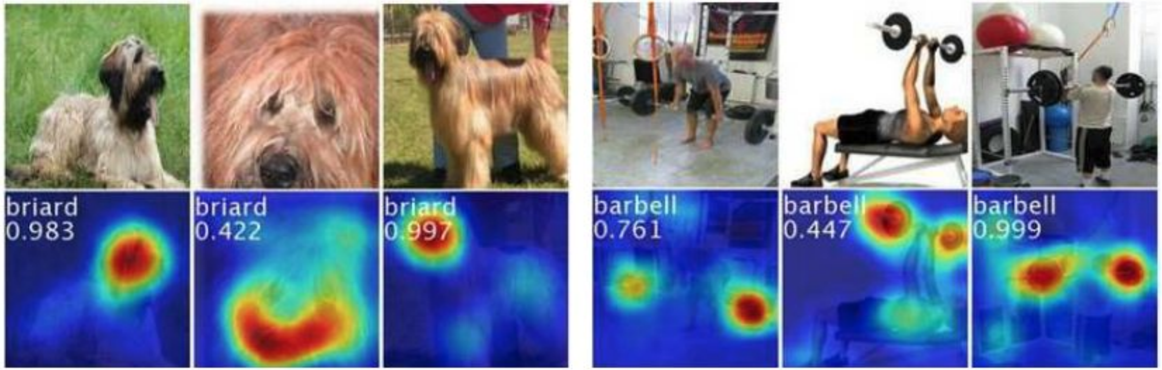
$$M_c(x, y) = \sum_k w_k^c f_k(x, y).$$

Weight (importance) of kth filter activation for predicting cth class

Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Saliency Maps: Class Activation Maps (CAM)

- Zhou et al. 2015
- Visualizes heatmap (class activation map) indicating the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c.



Zhou et al. Learning Deep Features for Discriminative Localization, 2016.

Grad-CAM: Extension of CAM to broader CNN architectures

- No longer relies on architectures that have a “global average pooling layer” at the end

Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2017.

Grad-CAM: Extension of CAM to broader CNN architectures

- No longer relies on architectures that have a "global average pooling layer" at the end

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

Gradient of prediction for cth class with respect to feature map activations A^k of a convolutional layer

Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2017.

Grad-CAM: Extension of CAM to broader CNN architectures

- No longer relies on architectures that have a “global average pooling layer” at the end

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

Weight (importance) of kth neuron in the CNN layer, for predicting the cth class

Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2017.

Grad-CAM: Extension of CAM to broader CNN architectures

- No longer relies on architectures that have a "global average pooling layer" at the end

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

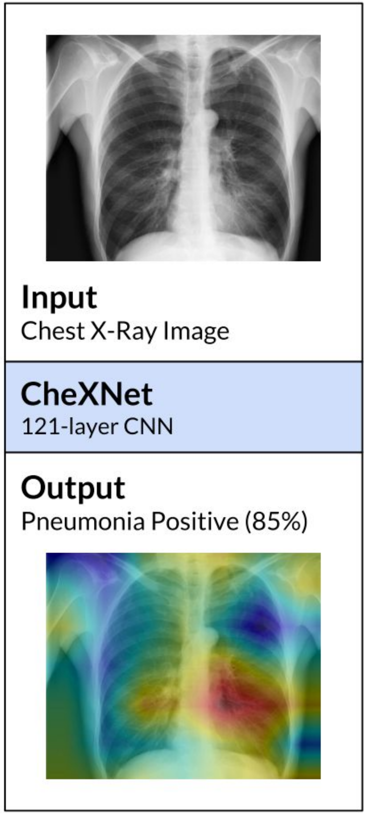
$$L^c_{\text{Grad-CAM}} = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k \right)}_{\text{linear combination}}$$

"Saliency" heatmap for the cth class is based on weighting a layer's activation map for each neuron by the importance of that neuron for predicting the class

Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2017.

Rajpurkar et al. 2017

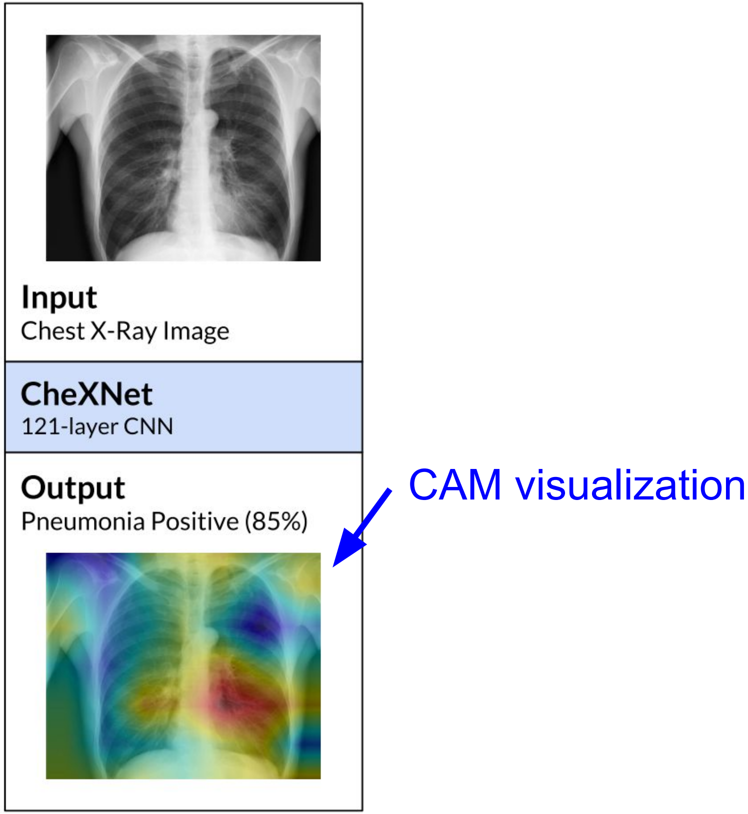
- Binary classification of pneumonia presence in chest X-rays
- Used ChestX-ray14 dataset with over 100,000 frontal X-ray images with 14 diseases
- 121-layer DenseNet CNN
- Compared algorithm performance with 4 radiologists
- Also applied algorithm to other diseases to surpass previous state-of-the-art on ChestX-ray14



Rajpurkar et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

Rajpurkar et al. 2017

- Binary classification of pneumonia presence in chest X-rays
- Used ChestX-ray14 dataset with over 100,000 frontal X-ray images with 14 diseases
- 121-layer DenseNet CNN
- Compared algorithm performance with 4 radiologists
- Also applied algorithm to other diseases to surpass previous state-of-the-art on ChestX-ray14



Rajpurkar et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

Rajpurkar et al. 2017

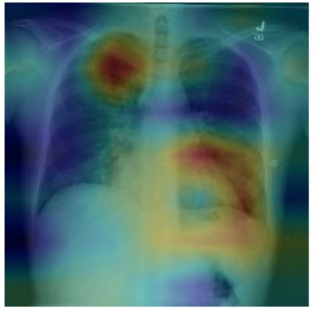
- Also showed CAM visualizations of predictions for other pathologies

Pathology	Wang et al. (2017)	Yao et al. (2017)	CheXNet (ours)
Atelectasis	0.716	0.772	**0.8094**
Cardiomegaly	0.807	0.904	**0.9248**
Effusion	0.784	0.859	**0.8638**
Infiltration	0.609	0.695	**0.7345**
Mass	0.706	0.792	**0.8676**
Nodule	0.671	0.717	**0.7802**
Pneumonia	0.633	0.713	**0.7680**
Pneumothorax	0.806	0.841	**0.8887**
Consolidation	0.708	0.788	**0.7901**
Edema	0.835	0.882	**0.8878**
Emphysema	0.815	0.829	**0.9371**
Fibrosis	0.769	0.767	**0.8047**
Pleural Thickening	0.708	0.765	**0.8062**
Hernia	0.767	0.914	**0.9164**

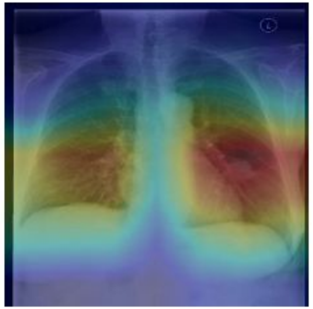
Rajpurkar et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

Rajpurkar et al. 2017

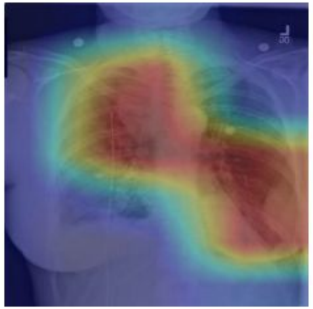
- Also showed CAM visualizations of predictions for other pathologies



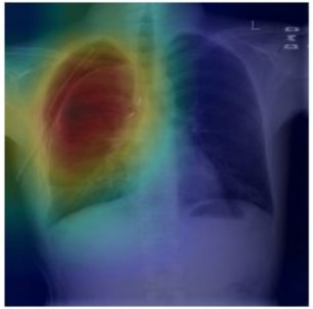
(a) Patient with multifocal community acquired pneumonia. The model correctly detects the airspace disease in the left lower and right upper lobes to arrive at the pneumonia diagnosis.



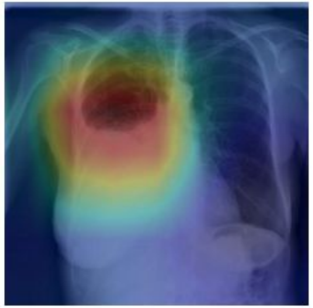
(b) Patient with a left lung nodule. The model identifies the left lower lobe lung nodule and correctly classifies the pathology.



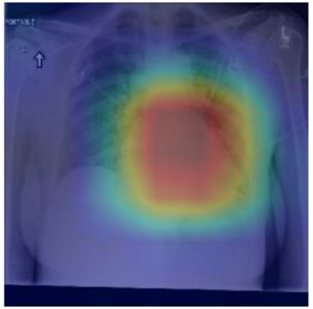
(c) Patient with primary lung malignancy and two large masses, one in the left lower lobe and one in the right upper lobe adjacent to the mediastinum. The model correctly identifies both masses in the X-ray.



(d) Patient with a right-sided pneumothroax and chest tube. The model detects the abnormal lung to correctly predict the presence of pneumothorax (collapsed lung).



(e) Patient with a large right pleural effusion (fluid in the pleural space). The model correctly labels the effusion and focuses on the right lower chest.

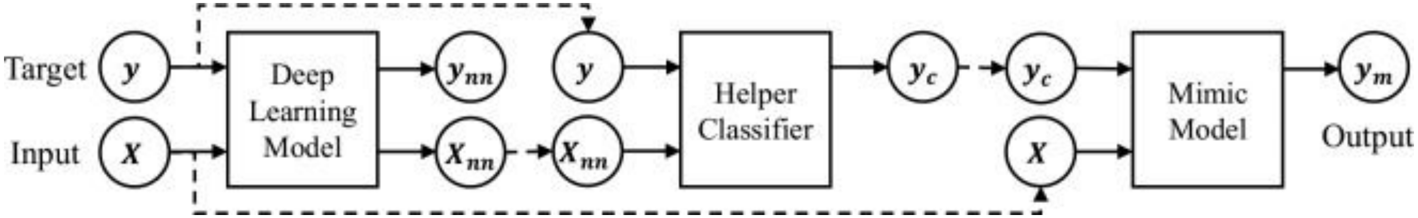


(f) Patient with congestive heart failure and cardiomegaly (enlarged heart). The model correctly identifies the enlarged cardiac silhouette.

Rajpurkar et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

A different type of approach: “distill” a complex neural network to an interpretable decision tree

- Che et al. 2017: distill deep neural network for ICU outcome prediction into an interpretable gradient boosting trees model (called mimic model)
- Benefits of distillation: 1) DNN can learn to correct for errors and noise in training data; 2) classification probabilities from DNN give “soft labels” containing more information; 3) Mimic approach can also be seen as a regularization on more complex DNN



Che et al. Interpretable Deep Models for ICU Outcome Prediction, 2016.

A different type of approach: “distill” a complex neural network to an interpretable decision tree

Methods		MOR (Mortality)		VFD (Ventilator Free Days)	
		AUROC	AUPRC	AUROC	AUPRC
Baselines	SVM	0.6437 ± 0.024	0.3408 ± 0.034	0.7251 ± 0.023	0.7901 ± 0.019
	LR	0.6915 ± 0.027	0.3736 ± 0.038	0.7592 ± 0.021	0.8142 ± 0.019
	DT	0.6024 ± 0.013	0.4369 ± 0.016	0.5794 ± 0.022	0.7570 ± 0.012
	GBT	0.7196 ± 0.023	0.4171 ± 0.040	0.7528 ± 0.017	0.8037 ± 0.018
Deep Models	DNN	0.7266 ± 0.089	0.4117 ± 0.122	0.7752 ± 0.054	0.8341 ± 0.042
	GRU	0.7666 ± 0.063	0.4587 ± 0.104	0.7723 ± 0.053	0.8131 ± 0.058
	DNN + GRU	0.7813 ± 0.028	**0.4874 ± 0.051**	**0.7896 ± 0.019**	**0.8397 ± 0.018**
Best Mimic Model		**0.7898 ± 0.030**	0.4766 ± 0.050	**0.7889 ± 0.018**	0.8324 ± 0.016

Che et al. Interpretable Deep Models for ICU Outcome Prediction, 2016.

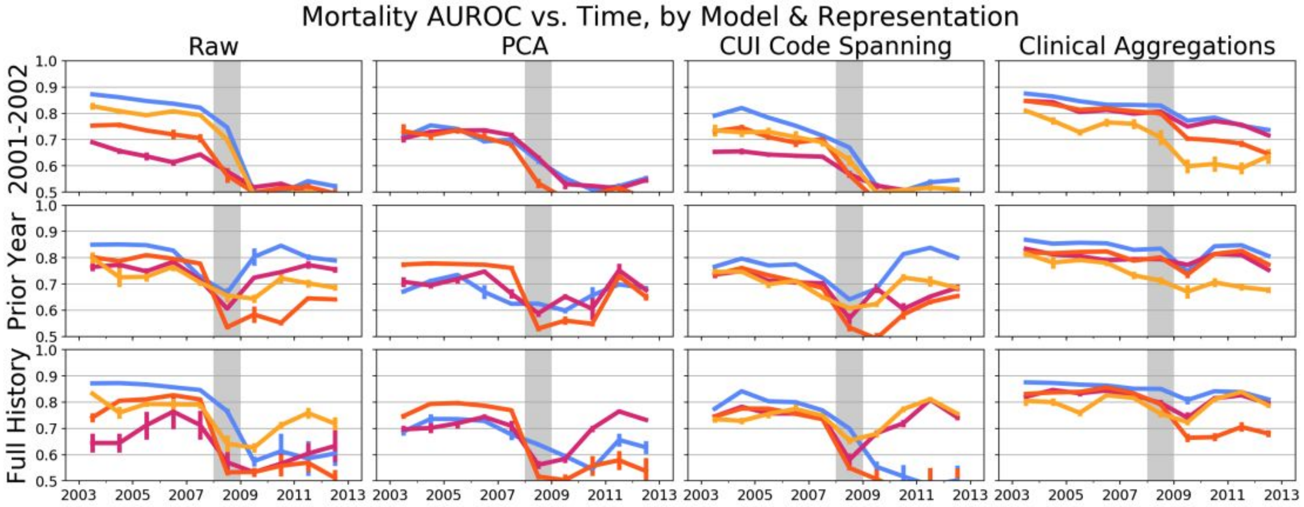
When can we trust the model?

- Notion of **uncertainty**: models can be more or less confident about a given prediction. Interpretability and explainability of the model gives indications of how the model arrived at its conclusion and how certain it is.

- Notion of **robustness**: models may behave differently under different settings (e.g. shift in the distribution of patient population / data). We may not be able to trust the model's outputs in the same way under some of these. Can we quantify how the model may perform under different settings, and make it "robust" under different settings that we care about?

Nestor et al. 2019

- Showed that EHR models using standard feature representations suffered drops in performance (evaluated by year) due to data drift from record keeping changes
- Introduced "clinical aggregations" of expert-defined similar clinical concepts for feature representations that increased robustness



Nestor et al. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks, 2019.

When can we trust the model?

- Notion of **uncertainty**: models can be more or less confident about a given prediction. Interpretability and explainability of the model gives indications of how the model arrived at its conclusion and how certain it is.

- Notion of **robustness**: models may behave differently under different settings (e.g. shift in the distribution of patient population / data). We may not be able to trust the model's outputs in the same way under some of these. Can we quantify how the model may perform under different settings, and make it "robust" under different settings that we care about?

Ideas like distributionally robust optimization minimize worst-case training loss over a set of groups (data distributions)

Ethics: many questions around AI / human collaboration in medicine

Ethics: many questions around AI / human collaboration in medicine

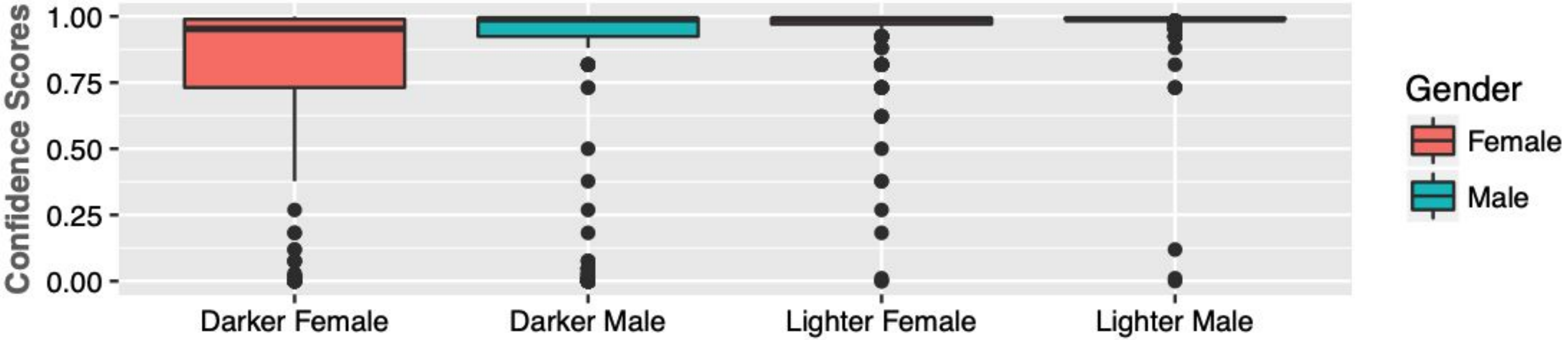
- How to make diagnosis and/or care decisions when the algorithm disagrees with the human?
- How should AI algorithms work together with humans?
- How to handle machine error vs. human error?
- How to make sure AI algorithms don't (perhaps inadvertently) discriminate against certain populations?
- How to handle tradeoffs between algorithmic performance on some groups vs. others?

Ethics: many questions around AI / human collaboration in medicine

- How to make diagnosis and/or care decisions when the algorithm disagrees with the human?
- How should AI algorithms work together with humans?
- How to handle machine error vs. human error?
- How to make sure AI algorithms don’t (perhaps inadvertently) discriminate against certain populations?
- How to handle tradeoffs between algorithmic performance on some groups vs. others?

Algorithmic bias

- Algorithm may perform better for one population vs. other, due to e.g. biases in training data or model
- E.g. Buolamwini and Gebru 2018: analysis of commercial gender classification systems by race

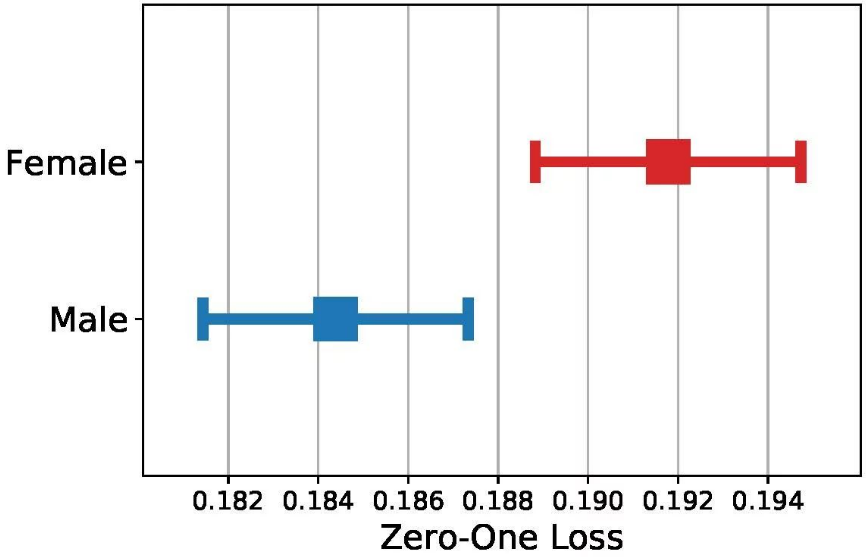


Buolamwini and Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, 2018.

Chen et al. 2019

- Showed discrepancies in error rates by race, gender, insurance type, etc. for models trained to make clinical predictions on MIMIC-III data

Error rate for predicting ICU mortality by gender

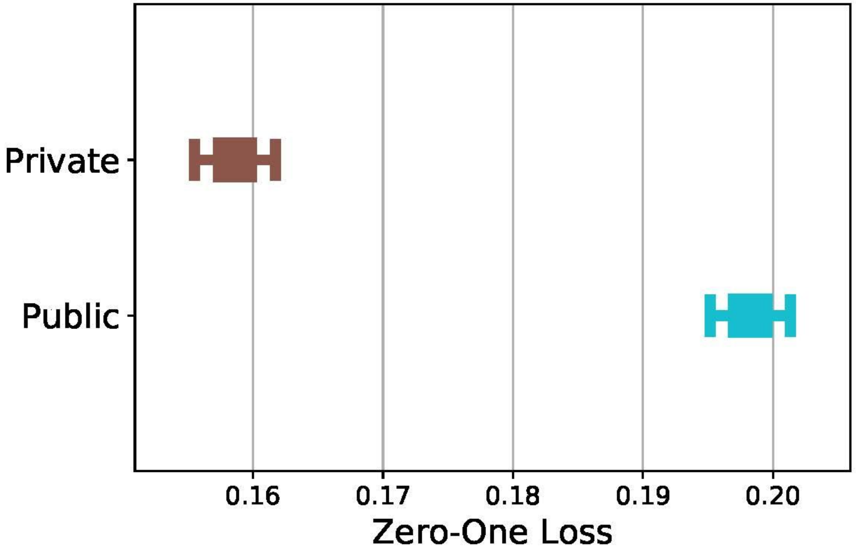


Chen et al. Can AI Help Reduce Disparities in General Medical and Mental Health Care? 2019.

Chen et al. 2019

- Showed discrepancies in error rates by race, gender, insurance type, etc. for models trained to make clinical predictions on MIMIC-III data

Error rate for predicting ICU mortality by insurance type

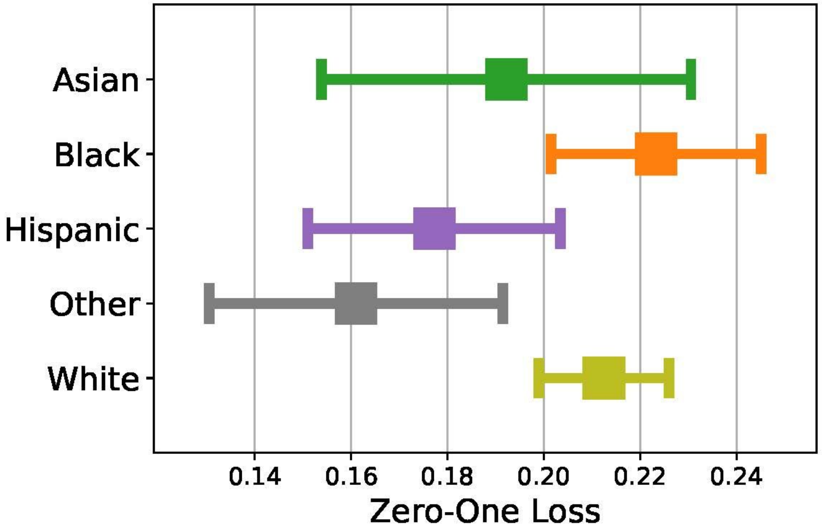


Chen et al. Can AI Help Reduce Disparities in General Medical and Mental Health Care? 2019.

Chen et al. 2019

- Showed discrepancies in error rates by race, gender, insurance type, etc. for models trained to make clinical predictions on MIMIC-III data

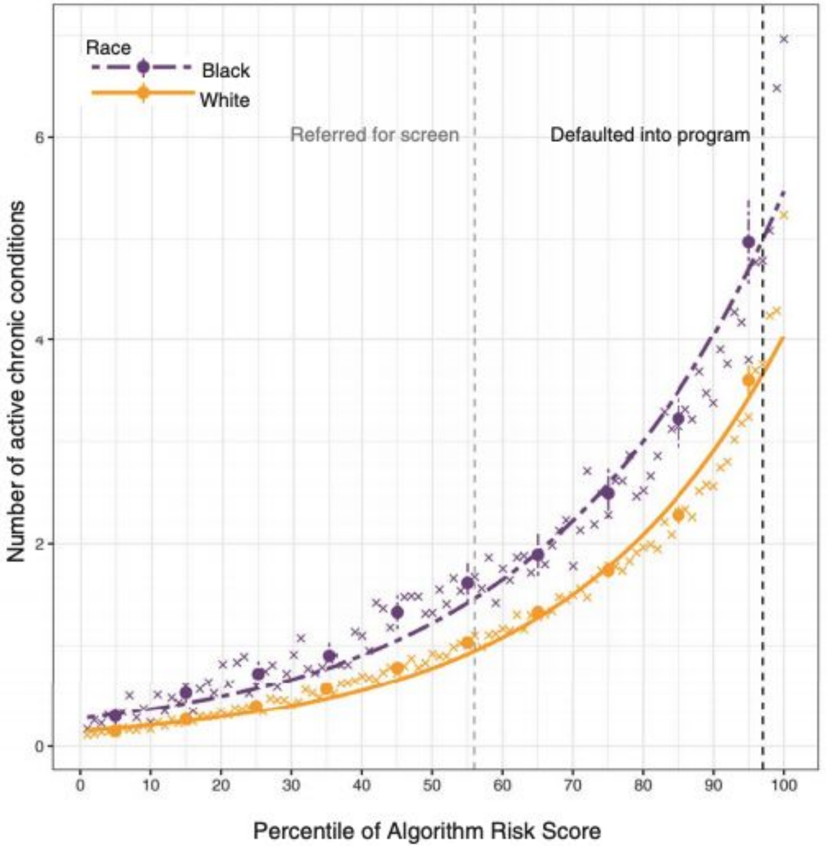
Error rate for predicting 30-day psychiatric readmission



Chen et al. Can AI Help Reduce Disparities in General Medical and Mental Health Care? 2019.

Obermeyer et al. 2019

- Finding that algorithm for allocating high-risk patients (complex medical needs) to special programs are less likely to refer black people vs. white people
- Algorithm used prediction of anticipated healthcare cost as a measure of complexity. But in training data, black patients had less healthcare cost for the same severity of sickness, due to less access to care
- Using other variables to predict risk reduced bias



Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations, 2019.

More on fairness… there are many possible definitions of fairness!

- **Group-independent predictions:** predictions should be independent of group membership
- **Equal metrics across groups**: e.g. equal true positive rates or false positive rates across groups
- **Individual fairness**: individuals who are similar with respect to a prediction task should have similar outcomes
- **Causal fairness**: e.g. there should not be a causal pathway from a sensitive attribute to the outcome prediction

Suresh and Guttag. A Framework for Understanding Unintended Consequences of Machine Learning, 2020.

More on fairness… there are many possible definitions of fairness!

- **Group-independent predictions:** predictions should be independent of group membership
- **Equal metrics across groups**: e.g. equal true positive rates or false positive rates across groups
- **Individual fairness**: individuals who are similar with respect to a prediction task should have similar outcomes
- **Causal fairness**: e.g. there should not be a causal pathway from a sensitive attribute to the outcome prediction

Cannot satisfy all of these simultaneously: satisfying "fairness" according to one definition generally leads to a trade-off respect to another definition!

Suresh and Guttag. A Framework for Understanding Unintended Consequences of Machine Learning, 2020.

Mitchell 2019: Model cards for Model Reporting

- Documentation accompanying trained models to detail performance characteristics

Model Card - Smiling Detection in Images

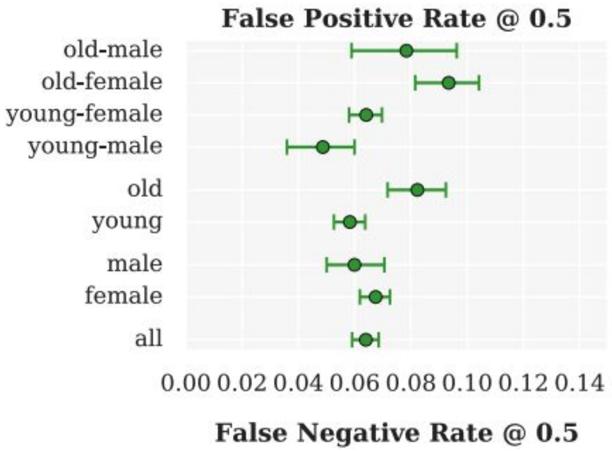
Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Quantitative Analyses



Mitchell et al. Model Cards for Model Reporting, 2019.

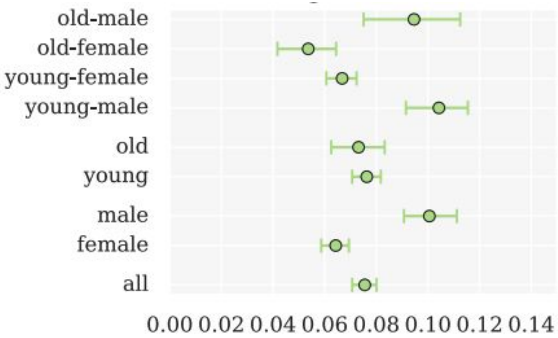
Mitchell 2019: Model cards for Model Reporting

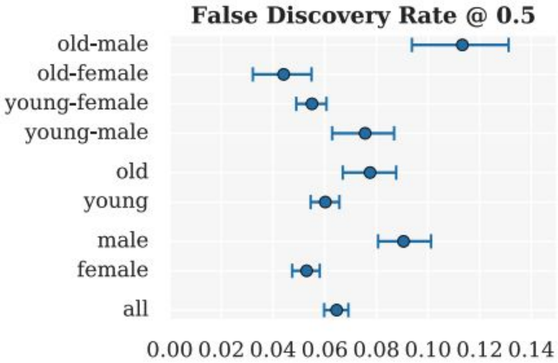
Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics corre-





Mitchell et al. Model Cards for Model Reporting, 2019.

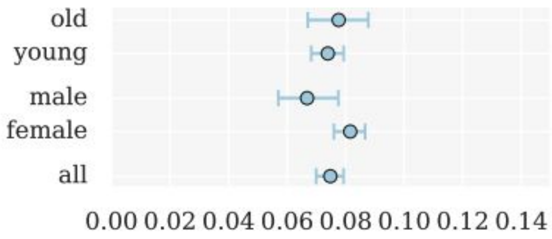
Mitchell 2019: Model cards for Model Reporting

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.



Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Mitchell et al. Model Cards for Model Reporting, 2019.

Gebru 2020: Datasheets for Datasets

A Database for Studying Face Recognition in Unconstrained Environments **Labeled Faces in the Wild**

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Gebru et al. Datasheets for Datasets. 2020.

Next time

- Distributed computing, security, and privacy