

# **Topic 4: Trustworthy Vision**

# Some Factors Affecting Trust in Deep Learning

- ▶ **Models Complexity - Non-Decomposability into Simple Components**
  - ▶ Explainability
  - ▶ Interpretability
- ▶ **Social Discrimination and Data/Model Misrepresentations**
  - ▶ Disparate Treatment (e.g. Social Biases in Datasets)
  - ▶ Disparate Impact (e.g. Discriminative Outcomes)
- ▶ **Unreliable Inference even to Minor Input Disruptions**
  - ▶ Adversarial Examples

# Impact of Stakeholders on Explainable AI (XAI)

## How do diverse stakeholders perceive about neural networks?

- ▶ Decision Maker
  - ▶ Use predictions as recommendations to make appropriate judgements
  - ▶ e.g. doctors trying to diagnose patients
  - ▶ Cares about global explanations as well as local explanations
- ▶ Affected User
  - ▶ Analyze their inputs in retrospect to change the future outcome
  - ▶ e.g. patients
  - ▶ Cares only about local explanations
- ▶ Regulator
  - ▶ Ensures the model is safe and compliant with
  - ▶ e.g. government official trying to validate the model
  - ▶ Cares about both global explanations and local explanations
- ▶ Data Scientist
  - ▶ Improve model performance
  - ▶ e.g. some of you in the future!

# Types of Explainable AI (XAI)

**Local Explanations:** Explain predictions for a given input data point

- ▶ Saliency Maps
- ▶ Class Activation Maps (CAM)
- ▶ Grad-CAM

**Global Explanations:** Explain the overall model

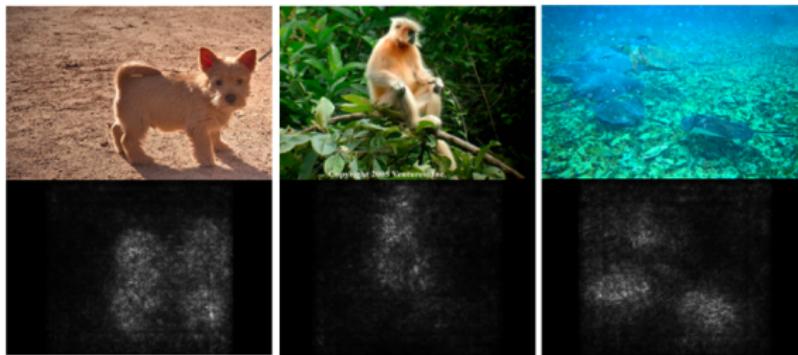
- ▶ ?

# Saliency Maps<sup>1</sup>

- ▶ Consider input image  $I_0$  of size  $m \times n$ , and a class  $c$
- ▶ Highly non-linear class score function  $S_c(I)$  in deep NNs  $\Rightarrow$   
Approximate  $S_c(I)$  with a linear function in the neighborhood of  $I_0$  using Taylor's expansion:

$$S_c(I) \approx w^T I + b, \text{ where } w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0} \text{ can be found via backprop.}$$

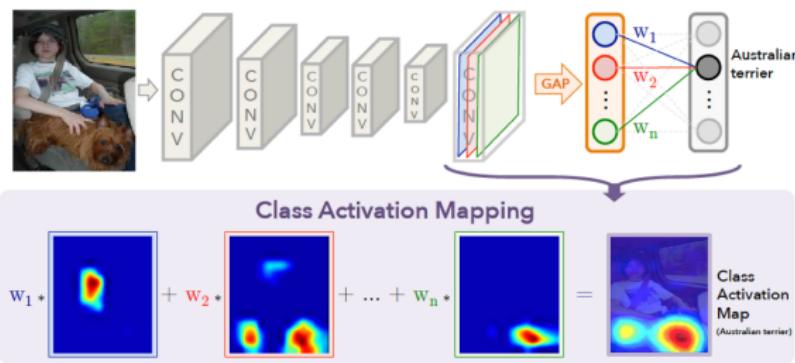
- ▶ Saliency Map  $M_{i,j} = |w_{h(i,j)}|$ , where  $h(i,j)$  is the index in  $w$  that corresponds to  $(i,j)^{th}$  pixel in  $I_0$ .
- ▶ Multi-channel images  $\Rightarrow M_{i,j} = \max_c |w_{h(i,j,c)}|$
- ▶ Also, a regression problem to produce images that maximize a given class score



<sup>1</sup> K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." ArXiv:1312.6034, 2013.

# Class Activation Maps (CAM<sup>2</sup>)

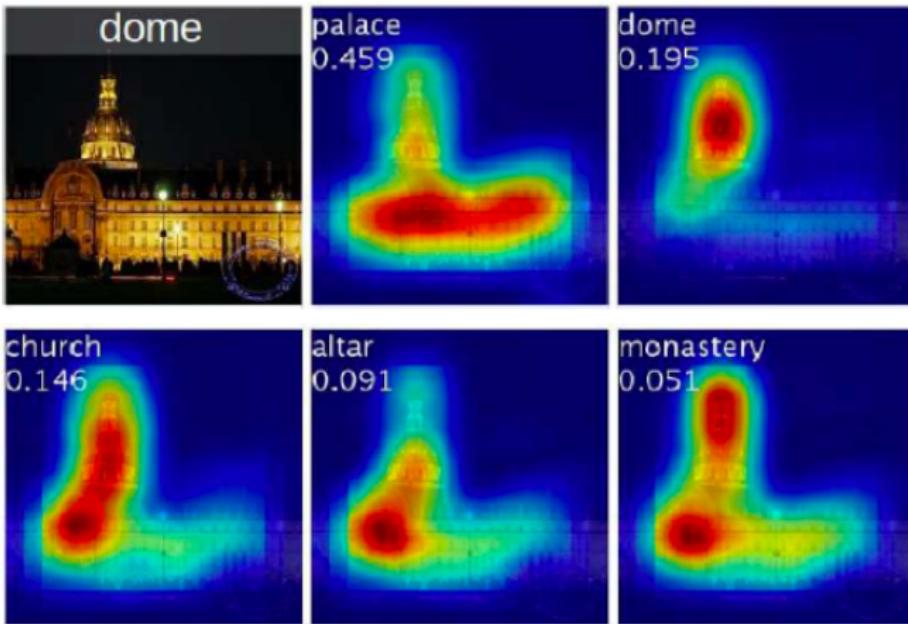
- ▶ Conv. layers are natural object detectors  $\Rightarrow$  Global average pooling (GAP) instead of FC layers
- ▶ Let  $f_k(x, y)$  denote activation of unit  $k$  at location  $(x, y)$ .
- ▶ Result of GAP at unit  $k$ :  $F_k = \frac{1}{Z} \sum_{x,y} f_k(x, y)$
- ▶ Class score:  $S_c = \sum_k w_k^c F_k$  (ignore bias term)  $\Rightarrow$  Softmax output:  $\frac{\exp(S_c)}{\sum_c \exp(S_c)}$
- ▶ CAM:  $M_c(x, y) = \sum_k w_k^c f_k(x, y) \Rightarrow S_c = \frac{1}{Z} \sum_{x,y} M_c(x, y)$
- ▶ Need to retrain the NN for weights  $w_k^c$
- ▶ Upscale CAM to input size.



<sup>2</sup>B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921-2929, 2016.

## CAM (cont...)

- ▶ Example of CAMs generated from top-5 predicted categories
- ▶ Note that the *dome* class activates the upper round portion, while *palace* activates the lower flat portion of the compound.



# Grad-CAM<sup>3</sup>

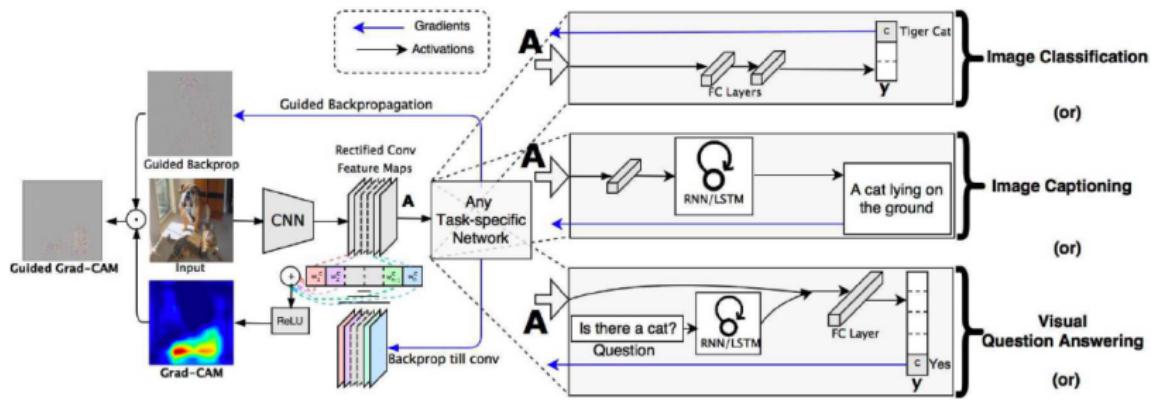
► Class score:  $S_c = \sum_k w_k^c F_k$ , where  $F_k = \frac{1}{Z} \sum_{x,y} f_k(x, y)$

► CAM:  $M_c(x, y) = \sum_k w_k^c f_k(x, y)$

►  $w_k^c = \frac{\partial S_c}{\partial F_k} = \frac{\partial S_c}{\partial f_k(x, y)} \left( \frac{\partial F_k}{\partial f_k(x, y)} \right)^{-1} = Z \cdot \frac{\partial S_c}{\partial f_k(x, y)}$

►  $\sum_{x,y} w_k^c = Z \cdot \sum_{x,y} \frac{\partial S_c}{\partial f_k(x, y)} \Rightarrow w_k^c = \sum_{x,y} \frac{\partial S_c}{\partial f_k(x, y)}$  (No need to retrain!)

► Grad-CAM:  $\tilde{M}_c(x, y) = \text{ReLU}[M_c(x, y)]$

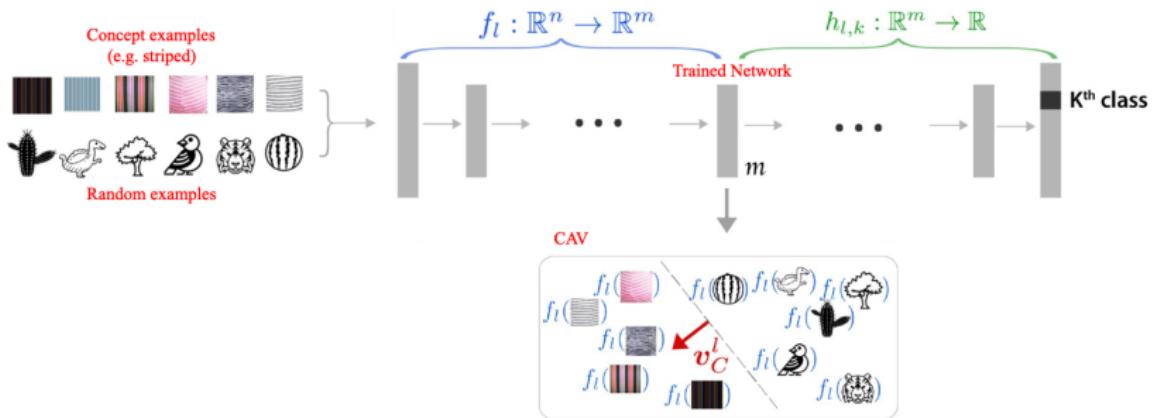


<sup>3</sup>R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626, 2017.

# Interpretable AI

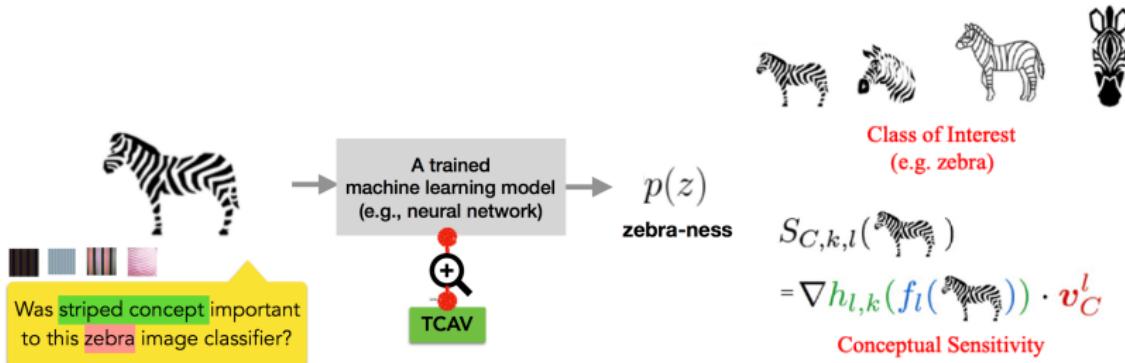
- ▶ Concept Activation Vectors (CAV)
- ▶ Uncertainty Quantification and Bayesian Neural Networks

# Concept Activation Vectors (CAV)



- ▶  $f_l(x)$  takes input  $x$  and outputs layer  $l$  activations  $a \in \mathbb{R}^M$ .
- ▶  $h_{l,k}(a)$  takes layer  $l$  activation  $a$  and outputs the class- $k$  logit  $\in \mathbb{R}$ .
- ▶ Given a user-defined concept  $C$ , let
  - ▶  $P_C$  denote the set of images that positively represent the concept
  - ▶  $N_C$  denote the set of images that negatively represent the concept
  - ▶  $A_P = \{f_l(x) | x \in P_C\}$ ,  $A_N = \{f_l(x) | x \in N_C\}$
- ▶ Train a linear classifier to find a hyperplane with normal  $v_C^l \in \mathbb{R}^M$  (CAV) that separates  $A_P$  and  $A_N$ .

# Testing with CAV (TCAV<sup>4</sup>)

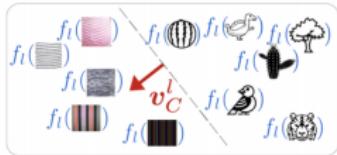


- ▶ **Conceptual Sensitivity:** A directional derivative  $S_{C,k,l}(x)$  that measures the sensitivity of logit output to change in CAV.
- ▶ In saliency maps, we compute the gradient wrt input pixels instead.
- ▶ **TCAV:** Aggregate per-input conceptual sensitivity over a class  $k$

$$TCAV_{C,k,l} = \frac{|\{x \in \mathcal{X}_k | S_{C,k,l}(x) > 0\}|}{|\mathcal{X}_k|}, \text{ where } \mathcal{X}_k \text{ denotes all inputs for class } k.$$

<sup>4</sup>B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viégas, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," In *International Conference on Machine Learning (ICML)*, pp. 2668-2677, 2018.

# Statistical Significance of CAV



Zebra

→  $\text{TCAV}_{Q_{C,k,l}}$  :

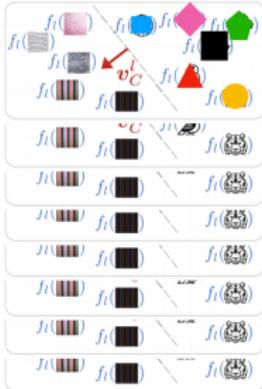
⋮

→  $\text{TCAV}_{Q_{C,k,l}}$  :

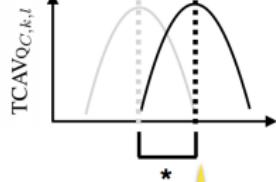
→  $\text{TCAV}_{Q_{C,k,l}}$  :

→  $\text{TCAV}_{Q_{C,k,l}}$  :

⋮



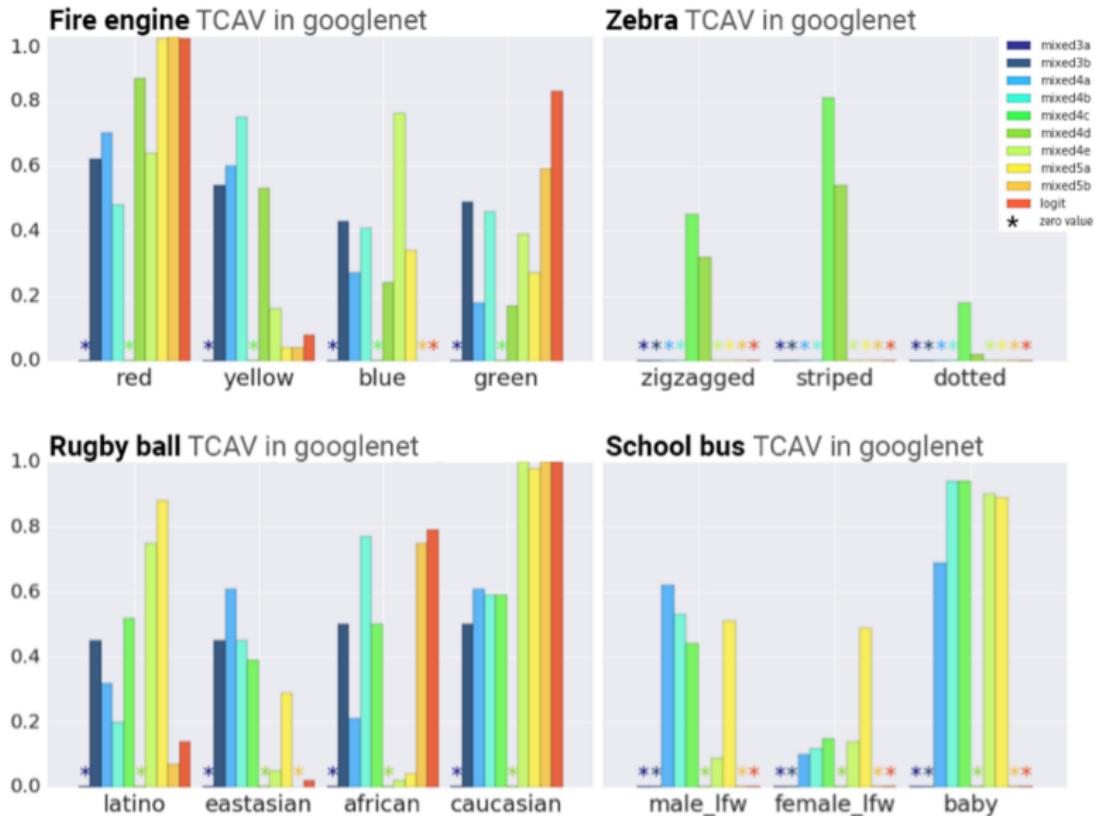
TCAV score  
random



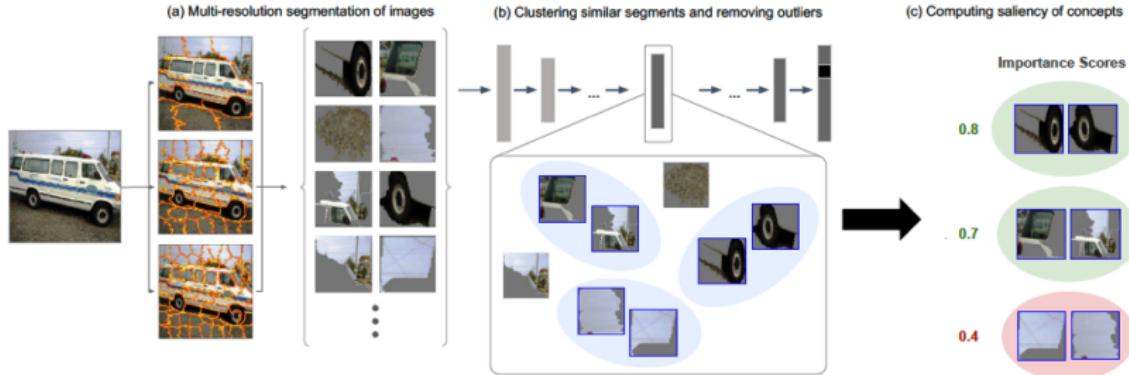
Check the distribution of  $\text{TCAV}_{Q_{C,k,l}}$  is statistically different from random using t-test

- ▶ Note: TCAV is very sensitive to low-quality random CAV.
- ▶ Compute TCAVs  $T$  times using different  $N_C$  sets to obtain  $\{\text{TCAV}_{Q_{C,k,l}}^{(i)}\}_{i=1}^T$
- ▶ Perform two-sided  $t$ -test.

# Example: TCAV on GoogLeNet



# Automatic Concept-Based Explanations (ACE<sup>5</sup>)



## Desired Properties of Concept-Based Explanation:

- ▶ **Meaningfulness:** Examples need to be semantically meaningful on its own. Also, multiple individuals should associate similar meaning to the same concept. (e.g. a group of pixels that contains a specific texture/object)
- ▶ **Coherency:** Examples need to be perceptually similar to each other, but also different from examples of other concepts.
- ▶ **Importance:** The concept's presence is necessary for the true prediction

<sup>5</sup> A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. "Towards Automatic Concept-Based Explanations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

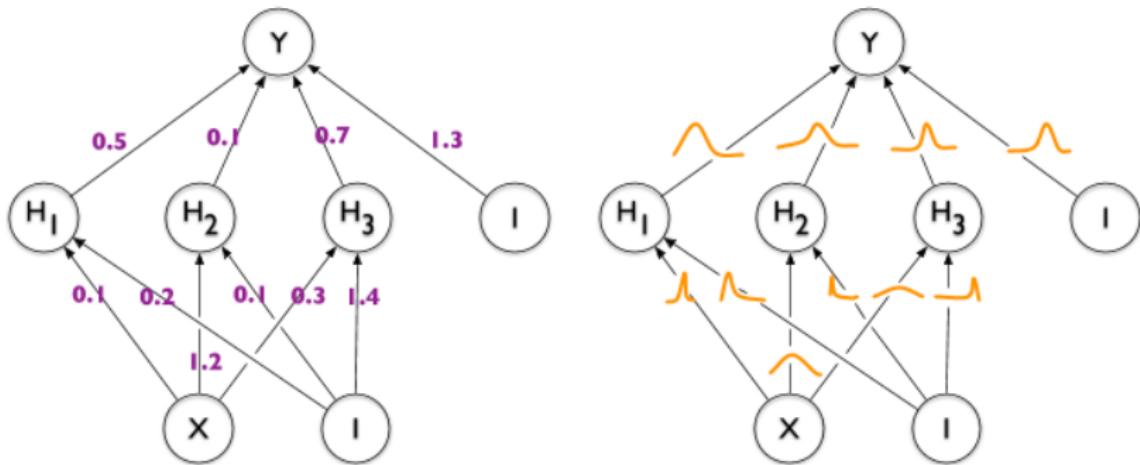
# Uncertainty Quantification

Two types of uncertainty:

- ▶ **Aleatoric Uncertainty:** Confidence in input data
  - ▶ High when input data is noisy
  - ▶ Cannot be reduced by adding more data
  - ▶ Can be estimated using likelihood methods using neural networks
- ▶ **Epistemic Uncertainty:** Confidence in Prediction
  - ▶ High when training data is small
  - ▶ Can be reduced by adding more data
  - ▶ Very difficult to estimate (Knowing when the model does not know the answer)

**Solution to Epistemic Uncertainty: Bayesian Neural Networks**

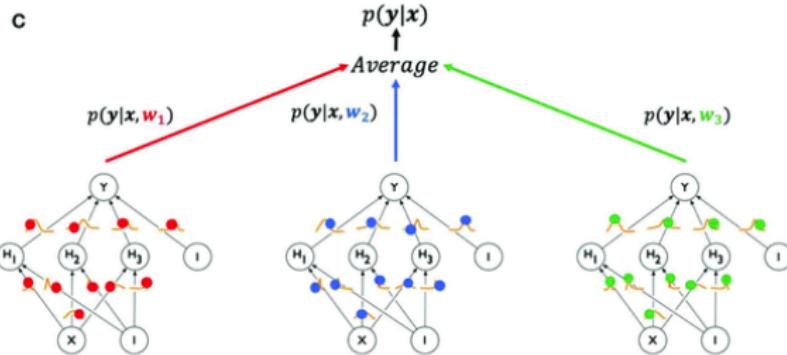
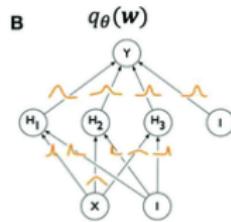
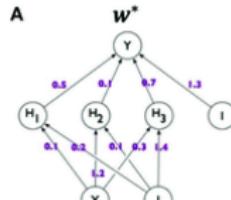
# Bayesian Neural Networks (BNNs)



- ▶ Train weight distributions, as opposed to just weights as in traditional NNs.
- ▶ Assume a prior distribution for weights  $p(\mathbb{W})$ , and a dataset  $(\mathbf{X}, \mathbf{Y})$ .
- ▶ Use Bayes' rule to update weight distribution via computing its posterior:

$$p(\mathbb{W}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X}, \mathbb{W}) \cdot p(\mathbb{W})}{p(\mathbf{Y}|\mathbf{X})}$$

# Emulating BNNs through Monte-Carlo Sampling<sup>7</sup>



- ▶ Sample weights from the trained distribution of weights several times
- ▶ Compute the average logit probability at the output of each class
- ▶ Similar approaches: Use Dropout<sup>6</sup> in Testing Phase to capture epistemic uncertainty

<sup>6</sup>Y. Gal, and Z. Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," In *International Conference on Machine Learning (ICML)*, pp. 1050-1059, PMLR, 2016.

<sup>7</sup>B. Lakshminarayanan, A. Pritzel, C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances of Neural Information Processing Systems (NeurIPS)*, 2017.

# ML Meets Social Sciences: Social Discrimination in CV<sup>8</sup>



- ▶ ML Designer's Perspective: Represent the data accurately
- ▶ Affected User's Perspective: **Is this algorithm fair? Am I (or, my group) being affected negatively by this solution?**
- ▶ Social biases observed in hindsight, after deployment
- ▶ Significant legal concern...
- ▶ Who is accountable for such flaws?

---

<sup>8</sup>J. Buolamwini, "The Coded Gaze: Unmasking Algorithmic Bias," YouTube: <https://www.youtube.com/watch?v=162VzSzzoPs>

# Racial Discrimination in Commercial Gender Classification<sup>9</sup>



Data Distribution vs. Skin Color Distribution

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	20.8	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	16.3	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	99.3	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	34.5	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	23.4	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	99.7
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	34.7	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	99.6	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	25.2	17.7	5.20	0.4



<sup>9</sup> J. Buolamwini, and T. Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," In Conference on Fairness, Accountability and Transparency (FAT), pp. 77-91, PMLR, 2018.

# Income-Based Discrimination in Object Detection<sup>10</sup>



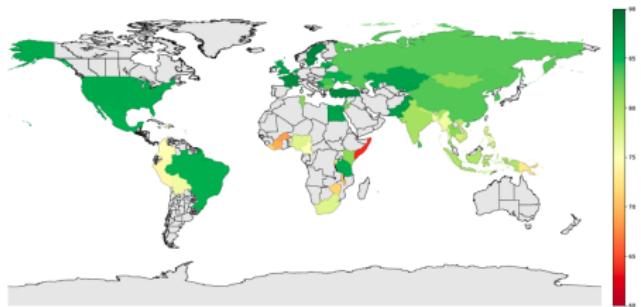
Ground truth: Soap      Nepal, 288 \$/month

Azure: food, cheese, bread, cake, sandwich  
Clarifai: food, wood, cooking, delicious, healthy  
Google: food, dish, cuisine, comfort food, spam  
Amazon: food, confectionary, sweets, burger  
Watson: food, food product, turmeric, seasoning  
Tencent: food, dish, matter, fast food, nutrient



Ground truth: Soap      UK, 1890 \$/month

Azure: toilet, design, art, sink  
Clarifai: people, faucet, healthcare, lavatory, wash closet  
Google: product, liquid, water, fluid, bathroom accessory  
Amazon: sink, indoors, bottle, sink faucet  
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser  
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave



<sup>10</sup>T. De Vries, I. Misra, C. Wang, and L. Van der Maaten, "Does Object Recognition Work for Everyone?," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 52-59. 2019.

# Biases in Pedestrian Detection

## Racial Discrimination in Pedestrian Detection<sup>11</sup>

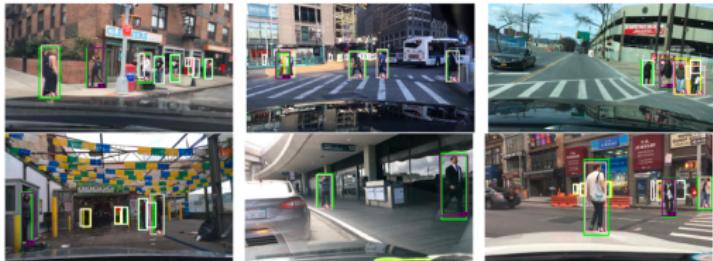
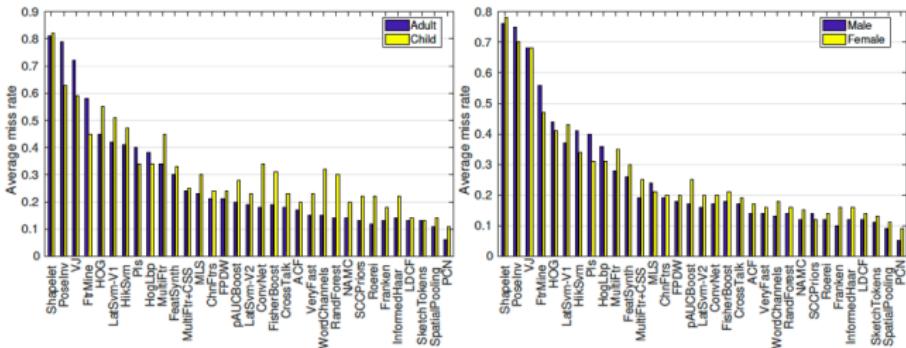


Figure 5. Example detections from Faster R-CNN using the R-50-FPN backbone, trained on BDD100K. For reference, the ground truth annotations for LS and DS are pink and purple respectively. Yellow boxes correspond to true positives under the AP<sub>50</sub> metric and false positives under the AP<sub>75</sub> metric. Green boxes correspond to true positives under the AP<sub>75</sub> metric. All the predictions shown are greater than an 85% confidence threshold.

Table 5. Average precision on BDD100K validation set with occluded individuals removed for models trained using MS COCO.

## Age/Gender Based Discrimination in Pedestrian Detection<sup>12</sup>

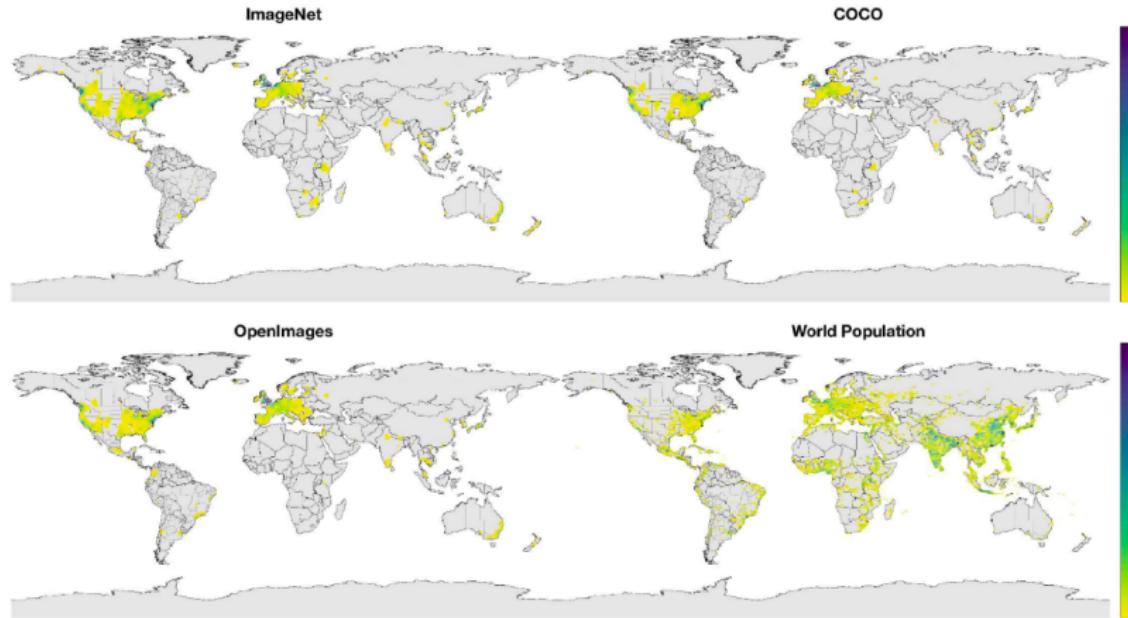


<sup>11</sup> B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive Inequity in Object Detection," ArXiv:1902.11097, 2019.

<sup>12</sup> M. Brandao, "Age and Gender Bias in Pedestrian Detection Algorithms," in *Workshop on Fairness Accountability Transparency and Ethics in Computer Vision (FATE/CV) in CVPR-2019*, Available: ArXiv: 1906.10490.

# What Causes Social Discrimination? <sup>13</sup>

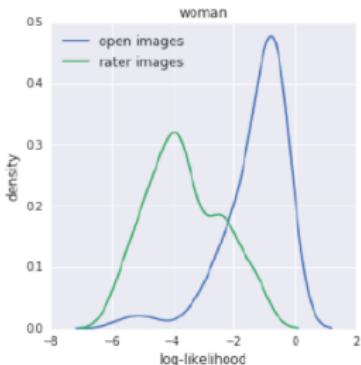
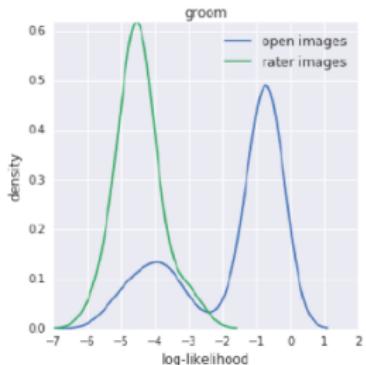
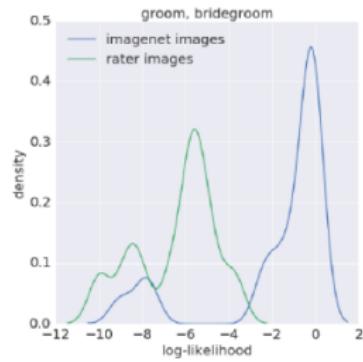
## Reason 1: Imbalanced Data Collection



<sup>13</sup>T. De Vries, I. Misra, C. Wang, and L. Van der Maaten, "Does Object Recognition Work for Everyone?", In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 52-59. 2019.

# What Causes Social Discrimination? <sup>14</sup>

## Reason 2: Biased Selection of Crowd Labelers



<sup>14</sup>T. De Vries, I. Misra, C. Wang, and L. Van der Maaten, "Does Object Recognition Work for Everyone?", In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 52-59. 2019.

# Evaluating Discrimination using Fairness Notions<sup>15</sup>

**Sources:** Systemic vs. Statistical

- ▶ **Systemic Discrimination:** Preference towards affected users who are similar to the decision maker. (e.g. job hiring)
- ▶ **Statistical Discrimination:** Using average group statistics to judge an individual (e.g. kidney matching)

**Types of Discrimination:** Direct vs. Indirect

- ▶ **Direct Discrimination:** Use of protected attributes of individuals to explicitly result in non-favorable outcomes (e.g. Admission decisions based on gender)
- ▶ **Indirect Discrimination:** Although seemingly neutral, protected groups still get to be treated unjustly due to implicit biases (e.g. gentrification and gerrymandering)

**Types of Discrimination:** Treatment vs. Impact

- ▶ **Disparate Impact** ⇒ *Obtain different impact at outcomes for different groups*
  - ▶ Group Fairness
- ▶ **Disparate Treatment** ⇒ *Treat individuals differently*
  - ▶ Individual Fairness

---

<sup>15</sup>N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1-35, 2021.

# Group Fairness

## Setting:

- ▶ Data tuple:  $(X, Y)$
- ▶ Groups:  $\mathcal{A}_1, \dots, \mathcal{A}_K$
- ▶ System Prediction:  $\hat{Y}$
- ▶ **Statistical/Demographic Parity:** A predictor  $\hat{Y}$  satisfies statistical/demographic parity if

$$\mathbb{P}(\hat{Y}|\mathcal{A}_i) = \mathbb{P}(\hat{Y}|\mathcal{A}_j), \text{ for all } i \neq j.$$

- ▶ **Calibration:** A predictor  $\hat{Y}$  is said to be calibrated if

$$\mathbb{P}(Y|\mathcal{A}_i, \hat{Y} = y) = \mathbb{P}(Y|\mathcal{A}_j, \hat{Y} = y), \text{ for all } i \neq j \text{ and } y \in \mathcal{Y}.$$

- ▶ **Equalized Odds:** A predictor  $\hat{Y}$  satisfies equalized odds if

$$\mathbb{P}(\hat{Y}|\mathcal{A}_i, Y = y) = \mathbb{P}(\hat{Y}|\mathcal{A}_j, Y = y), \text{ for all } i \neq j \text{ and } y \in \mathcal{Y}.$$

# Individual Fairness

- ▶ **Fairness through Unawareness:** No protected attributes are explicitly used in the decision-making process
- ▶ **Fairness through Awareness:** Similar individuals with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome.
- ▶ **Counterfactual Fairness:** A decision is fair if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group. In other words,

$$\mathbb{P}(\hat{Y}_{\mathcal{A} \leftarrow \mathcal{A}_i} | X, \mathcal{A}) = \mathbb{P}(\hat{Y}_{\mathcal{A} \leftarrow \mathcal{A}_j} | X, \mathcal{A}), \text{ for all } i \neq j.$$

# Mitigating Discrimination

- ▶ Blinding
- ▶ Debiasing
- ▶ Pre-Processing
- ▶ In-Processing
- ▶ Post-Processing