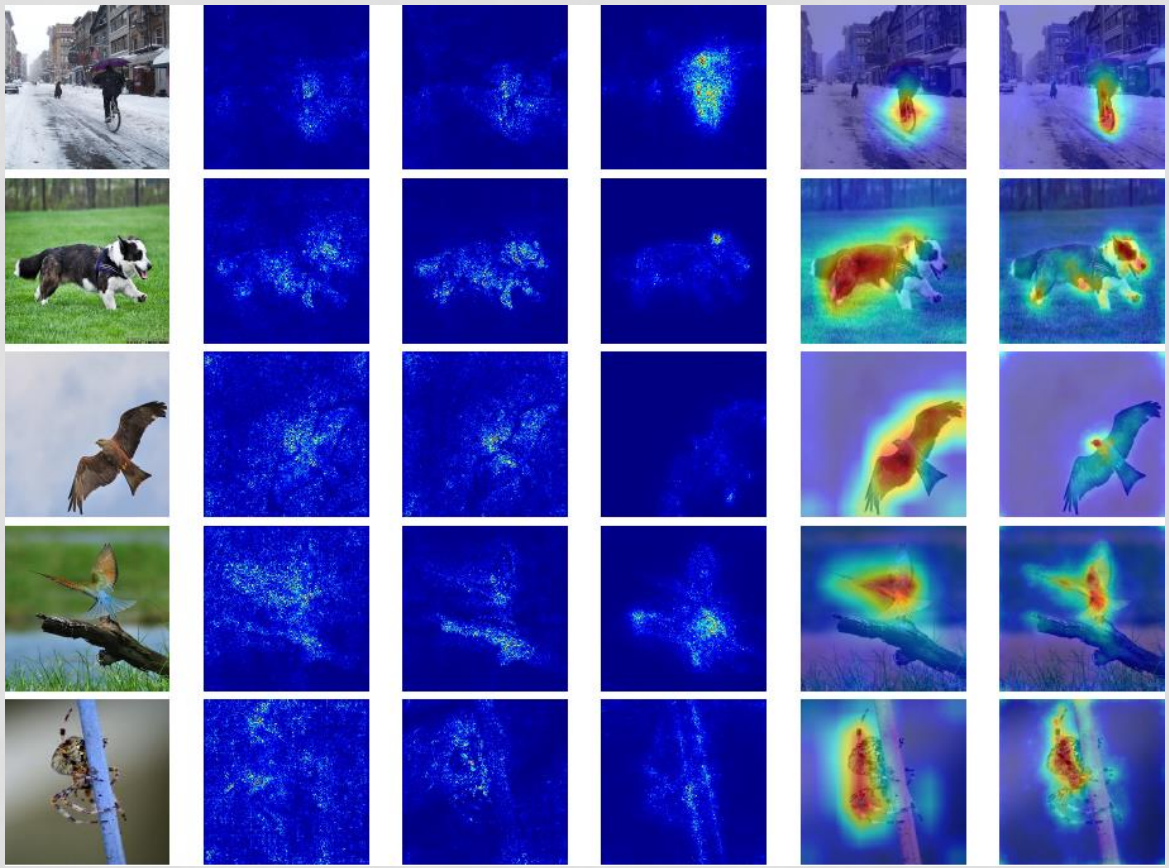
INTERPRETING NN RESPONSES

TRANSPARENCY & FAIRNESS

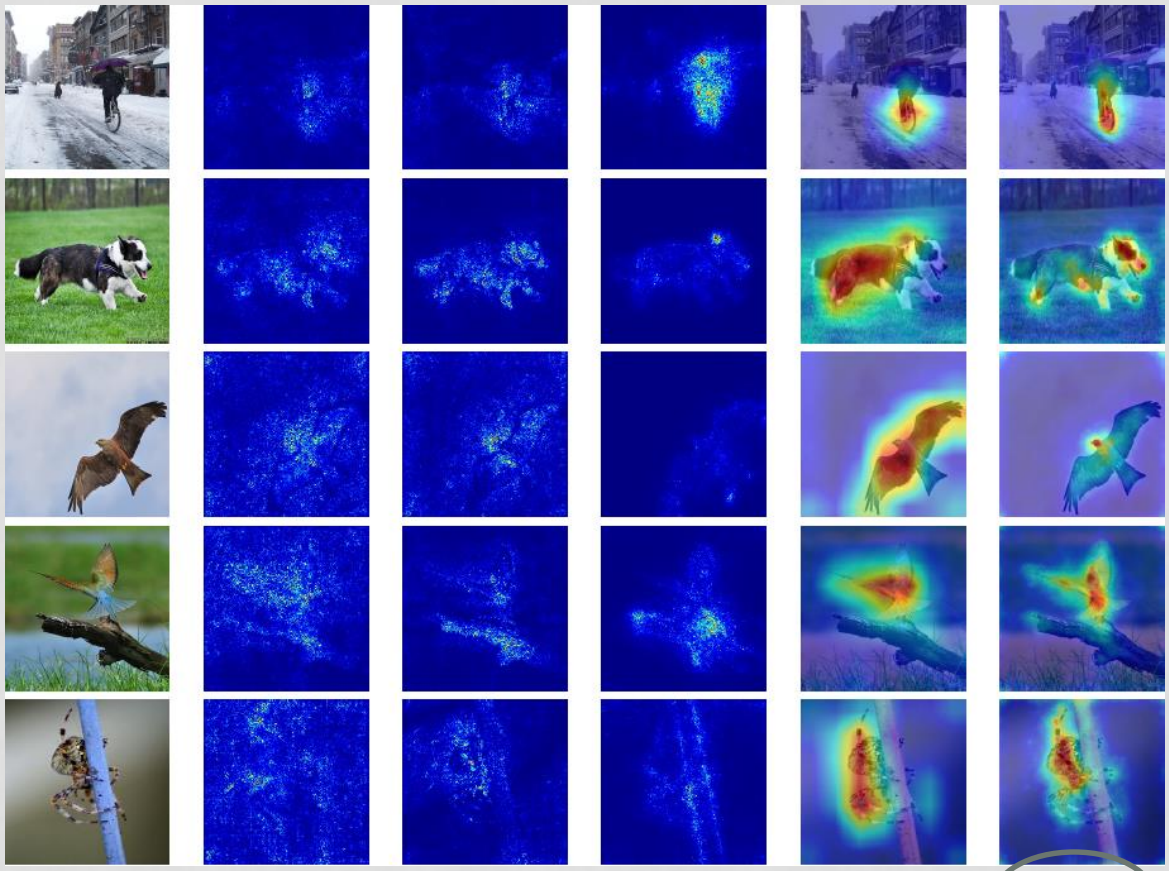
P.CHANDRIKASINGH, R. LEUSHUIS, P. LINTL & A. VICOL

NEW SALIENCY METHOD



Source: S. Srinivas and F. Fleuret. Full-Gradient Representation for Neural Network Visualization. In NeurIPS 2019, 2019

NEW SALIENCY METHOD



Source: S. Srinivas and F. Fleuret. Full-Gradient Representation for Neural Network Visualization. In NeurIPS 2019, 2019

CONTENTS

- How it works
- Performance
  - Pixel Perturbation (PP) framework
  - RemOve And Retrain (ROAR) framework
- Detecting bias
- Conclusion

HOW IT WORKS

- Full-gradients

$$G = (\nabla_x f(x, b),\ \nabla_b f(x, b) \odot b)$$

- Better than saliency maps
  - Weak dependence
  - Completeness

HOW IT WORKS

- Full-gradients

$$G = (\nabla_x f(x, b),\ \nabla_b f(x, b) \odot b)$$

- Reduce to saliency map → FullGrad

$$S_f(x) = \nabla_x f(x, b) \odot x + \sum_{l \in L} \sum_{c \in c_l} (\nabla_b f(x, b) \odot b)_c$$

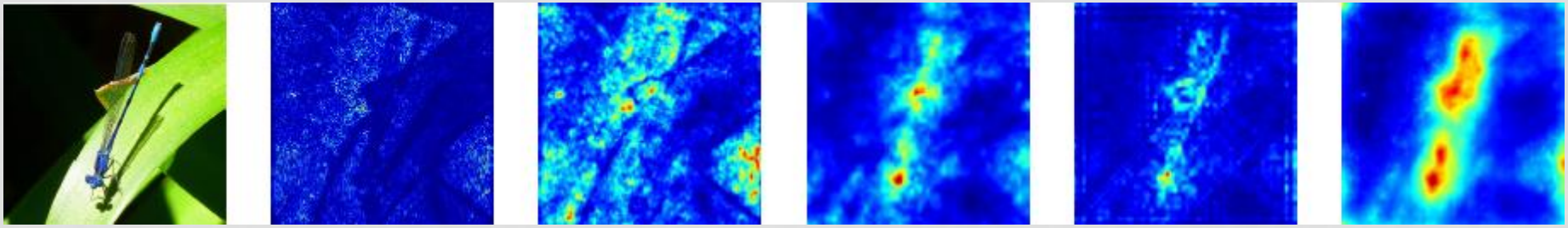


Image | Input-grad x input | Layer 3 bias-gradient | Layer 5 bias-gradient | Layer 7 bias-gradient | FullGrad aggregate

Source: S. Srinivas and F. Fleuret. Full-Gradient Representation for Neural Network Visualization. In NeurIPS 2019, 2019

HOW IT WORKS

- Reduce to saliency map

$$S_f(x) = \nabla_x f(x, b) \odot x + \sum_{l \in L} \sum_{c \in c_l} (\nabla_b f(x, b) \odot b)_c$$

- With post-processing operator ψ

$$S_f(x) = \psi(\nabla_x f(x, b) \odot x) + \sum_{l \in L} \sum_{c \in c_l} \psi\big((\nabla_b f(x, b) \odot b)_c\big)$$

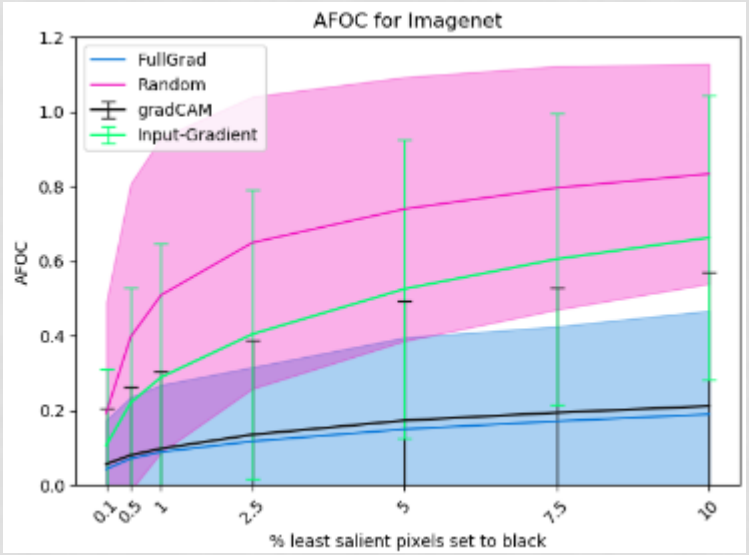
PERFORMANCE

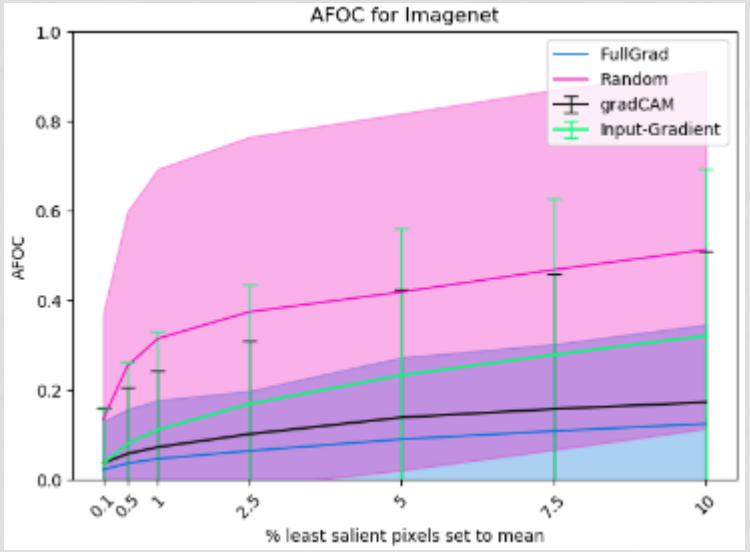
- Pixel Perturbation (PP) framework
- RemOve And Retrain (ROAR) framework

PP FRAMEWORK

- Original scheme
  - k% most salient
  - artifacts
- Augmented scheme
  - k% least salient
  - Unimportant regions
- Remove or set to mean

PP FRAMEWORK





- Absolute fractional output change

$$AFOC = \frac{|FFN(x_k) - FFN(x)|}{FNN(x)}$$

PP FRAMEWORK

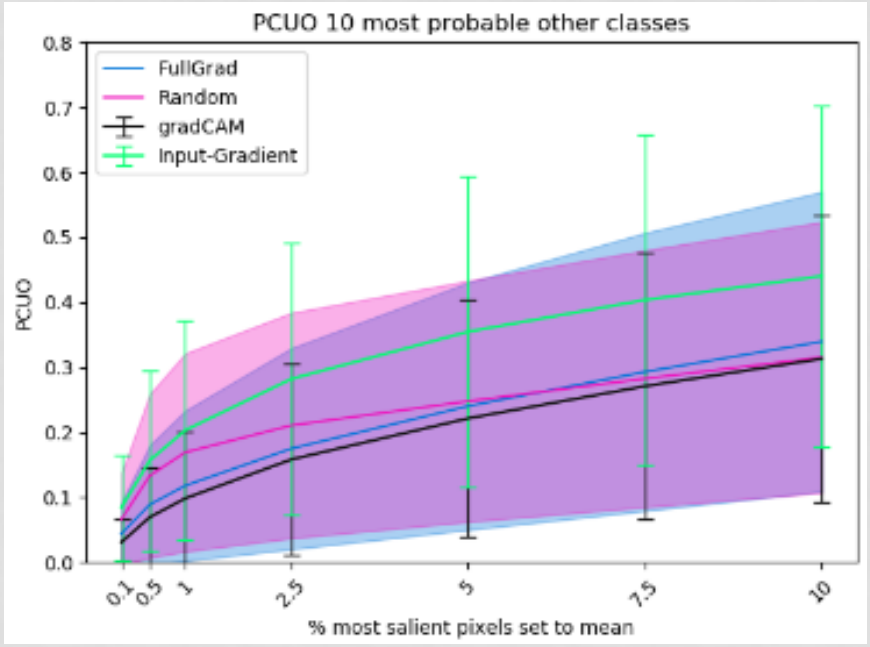
- KL divergence

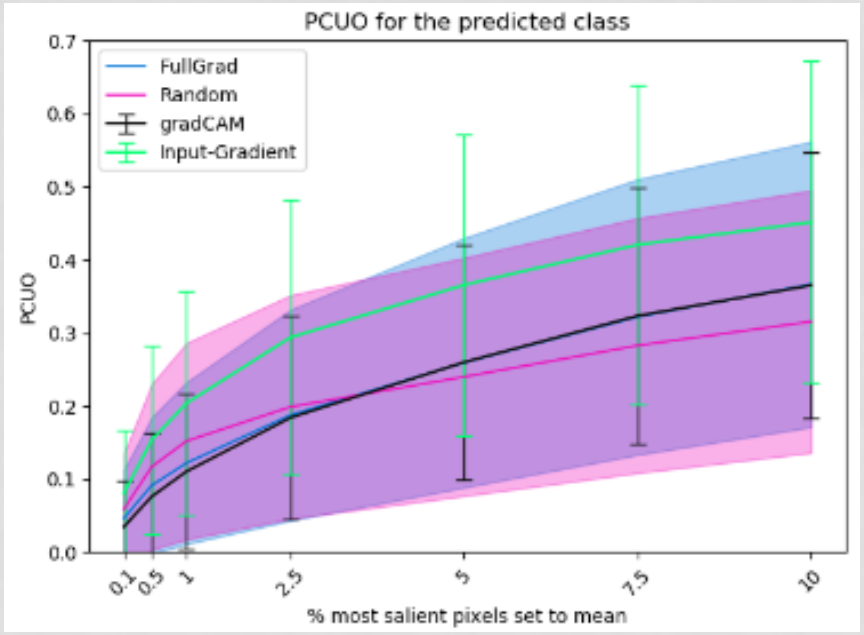
$$D_{KL}(P_k|Q) = \sum_{x \in X} P_k(x_k) \log\left(\frac{P_k(x_k)}{Q(x)}\right)$$

- Percentage change in the unnormalized output

$$PCUO = \frac{|FNN_i^{-1}(x_k) - FNN_i^{-1}(x)|}{FNN_i^{-1}(x)}$$

PP FRAMEWORK

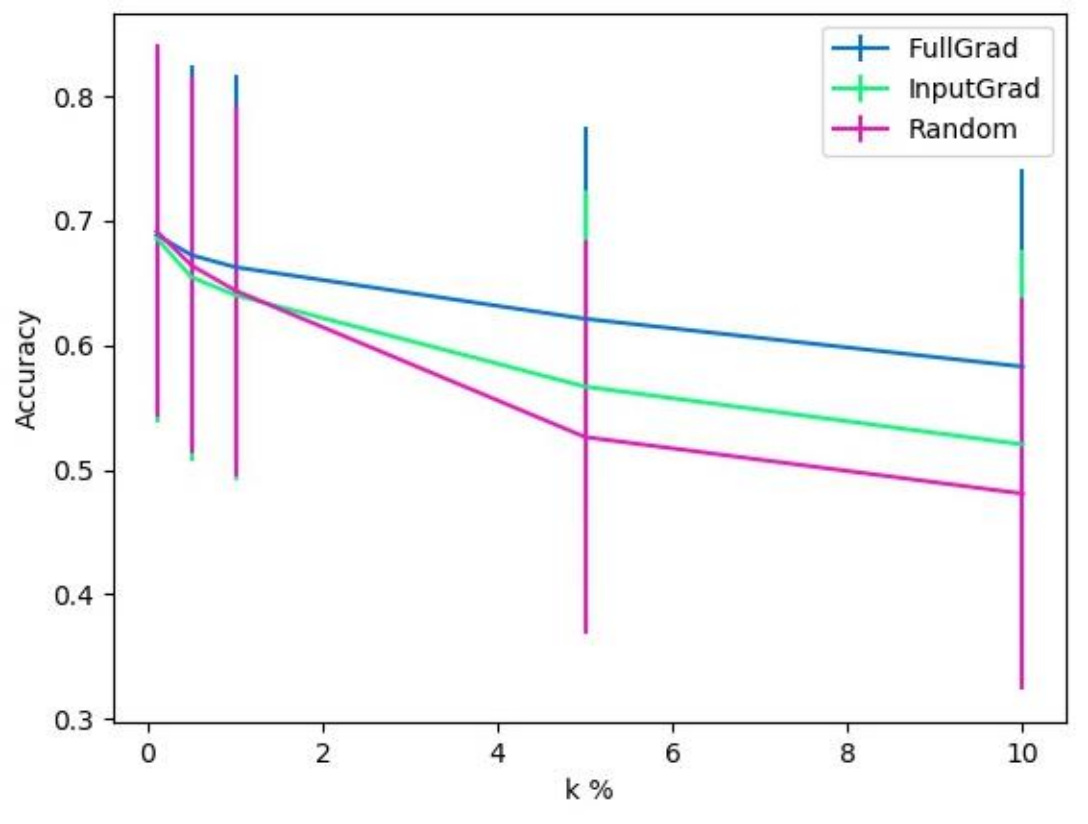




PERFORMANCE

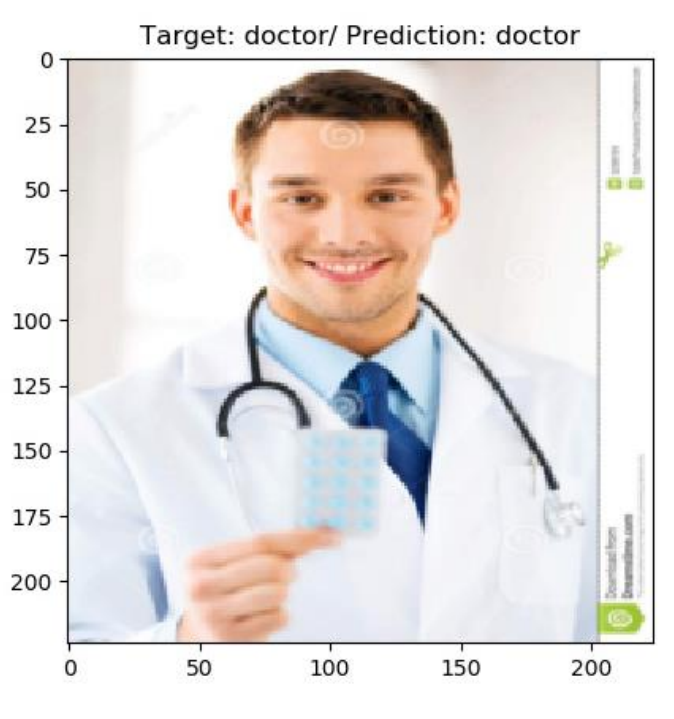
- Pixel Perturbation (PP) framework
- RemOve And Retrain (ROAR) framework

ROAR FRAMEWORK

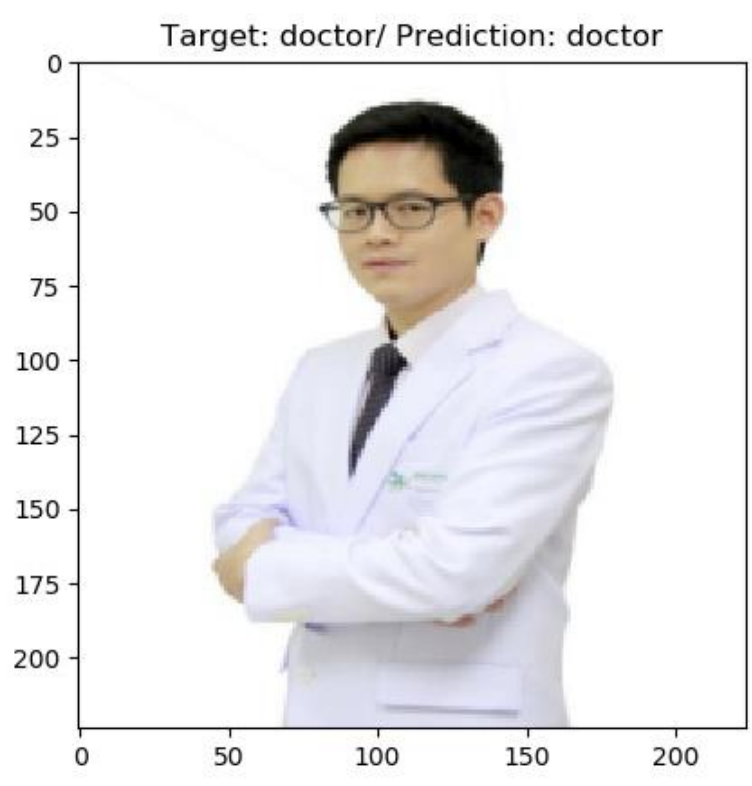


$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} I(predicted_i = target_i)$$

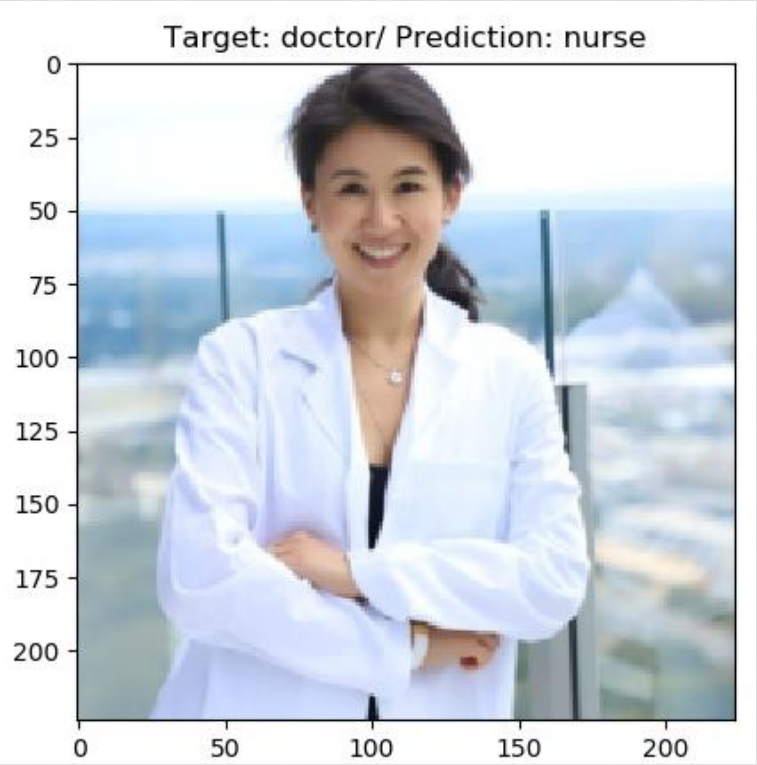
DETECTING BIAS



DETECTING BIAS



DETECTING BIAS



DETECTING BIAS

CONCLUSION

- Highly effective
- Not significant
- Fairness and accountability purposes
- Still a lot of research