# Explainable AI for Deep Learning

Andrew H. Fagg

# Explainable AI Methods

- Want to understand ***how*** our learned models make their decisions
- Hard due to the large number of parameters and interactions
  - A particular challenge with deep, non-sequential models

# XAI is Important

- Convince ourselves that the methods are learning meaningful and robust representations
- Guide further modeling approaches
- Domain scientists/engineers need to be able to relate the learned models to their knowledge
  - Otherwise, acceptance can be difficult
- Opportunity to discover new domain knowledge
- Verification that learned models are making appropriate decisions (including ethical ones)

Our focus today is on image classifiers…

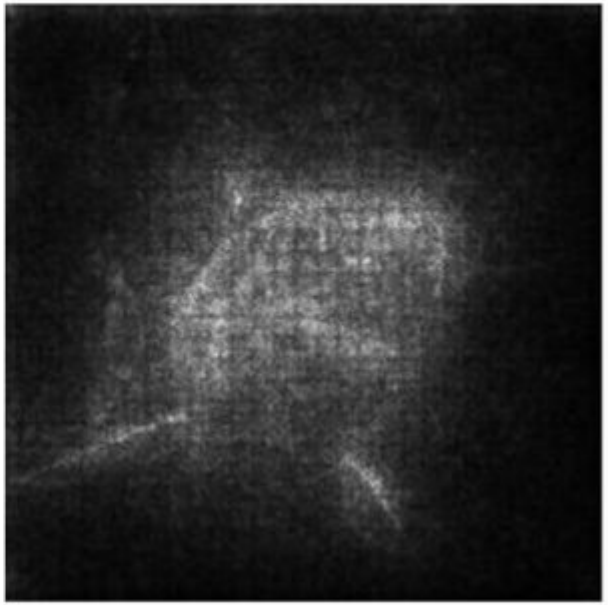
# Saliency Maps

What input pixels are most important for making the class label decision for a specific input image?

- Magnitude of d L / d I
  - L is class probability output
- For one image, this tensor is also image r x c
- Magnitude tells us which pixels contribute the most
- Sign tells us the direction of influence

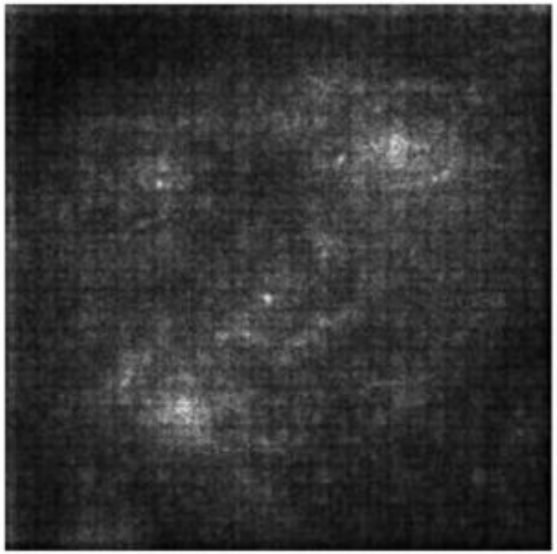
Bach et al. 2015

# Saliency Examples

| ImageNet original image (great white shark) | Corresponding saliency map (great white shark) |
| --- | --- |
|  |  |

https://medium.com

# Saliency Examples

| PubFig original image (Barack Obama) | Corresponding saliency map (Barack Obama) |
| --- | --- |
|  |  |

https://medium.com

# Saliency Concerns

Can be strangely specific

- Focus on pixels, not on regions of the image
- If some salient pixels are changed, would this change our probability in a substantial way?
- … other pixels may then increase in saliency, but essentially give the same prediction

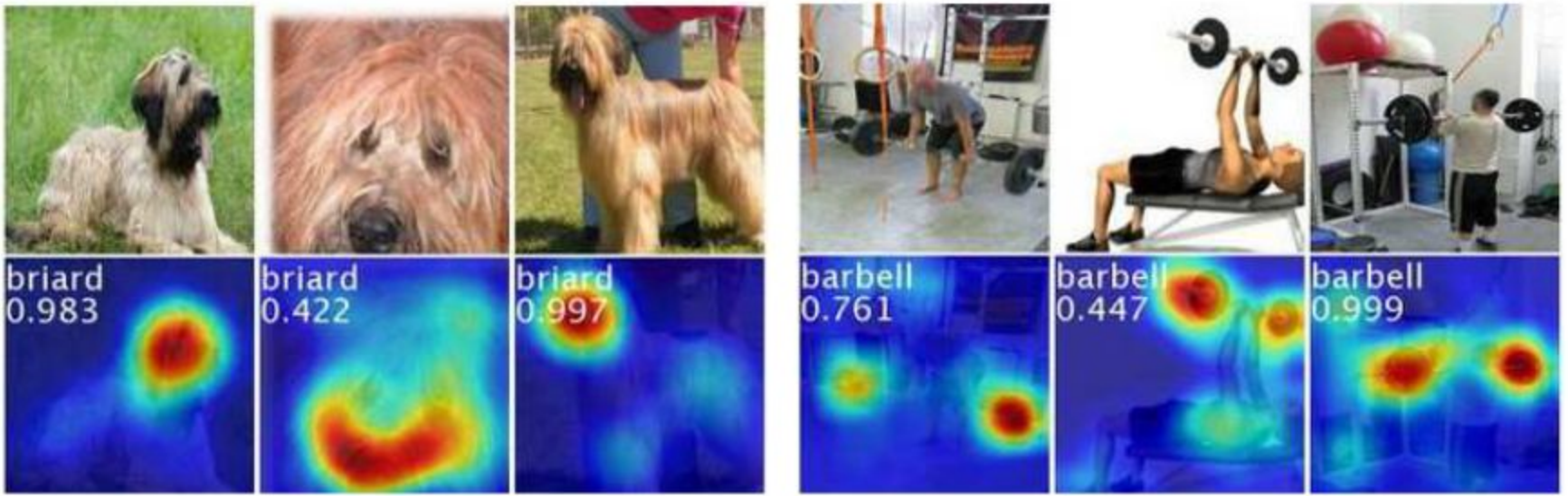
# Class Activation Map

Zhao et al. (2016) CVPR

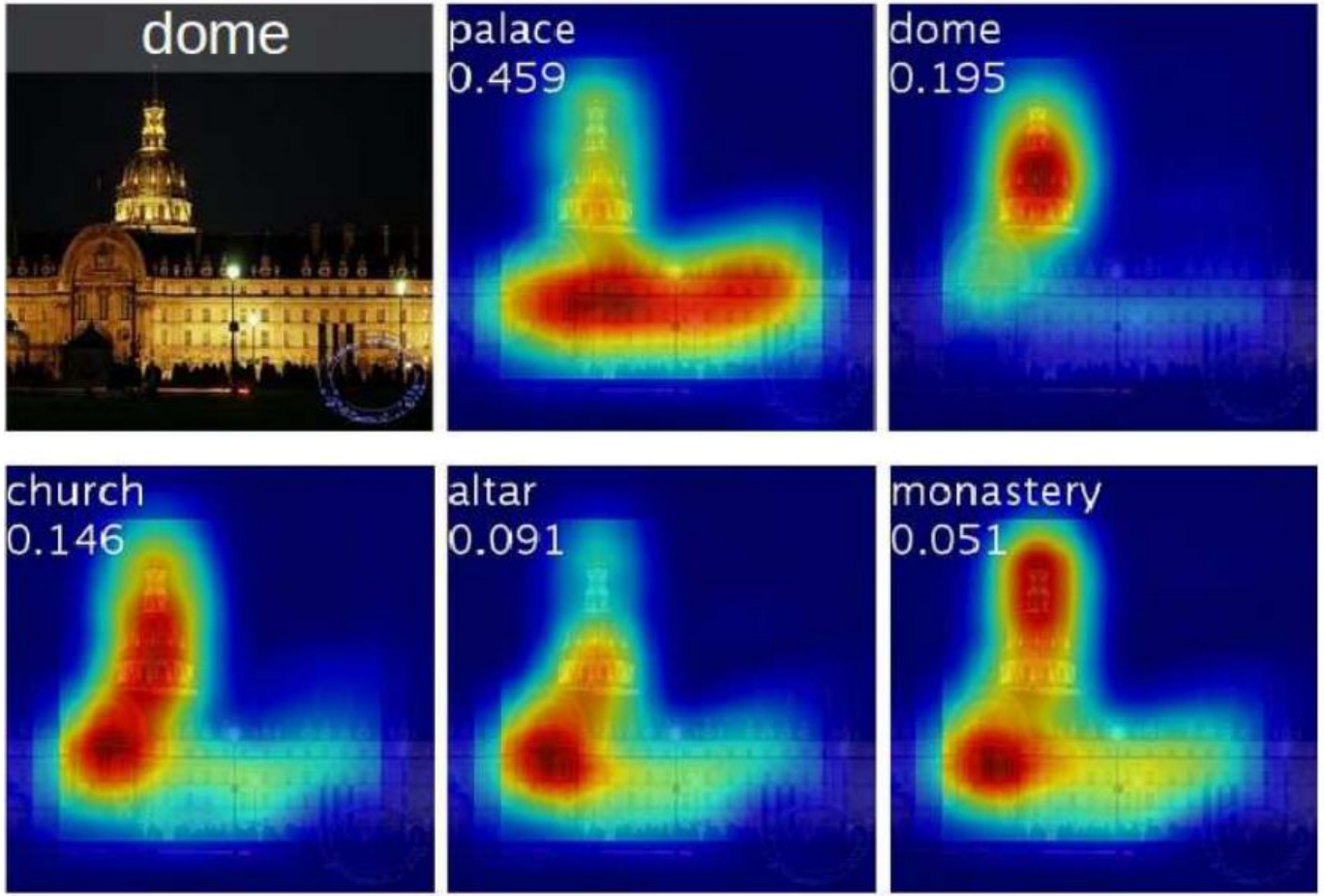
- Start with image classification architecture
- Focus on output of convolutional layers
  - Still have some spatial information
- This gives us a sense of the most important parts of the image for a given class

# CAM Visualization

- Scale CAM map back to original resolution
- Smooth with Gaussian filter
- (in some cases) overlay the original image




Zhao et al. (2016)




Zhao et al. (2016)

# CAM Challenges

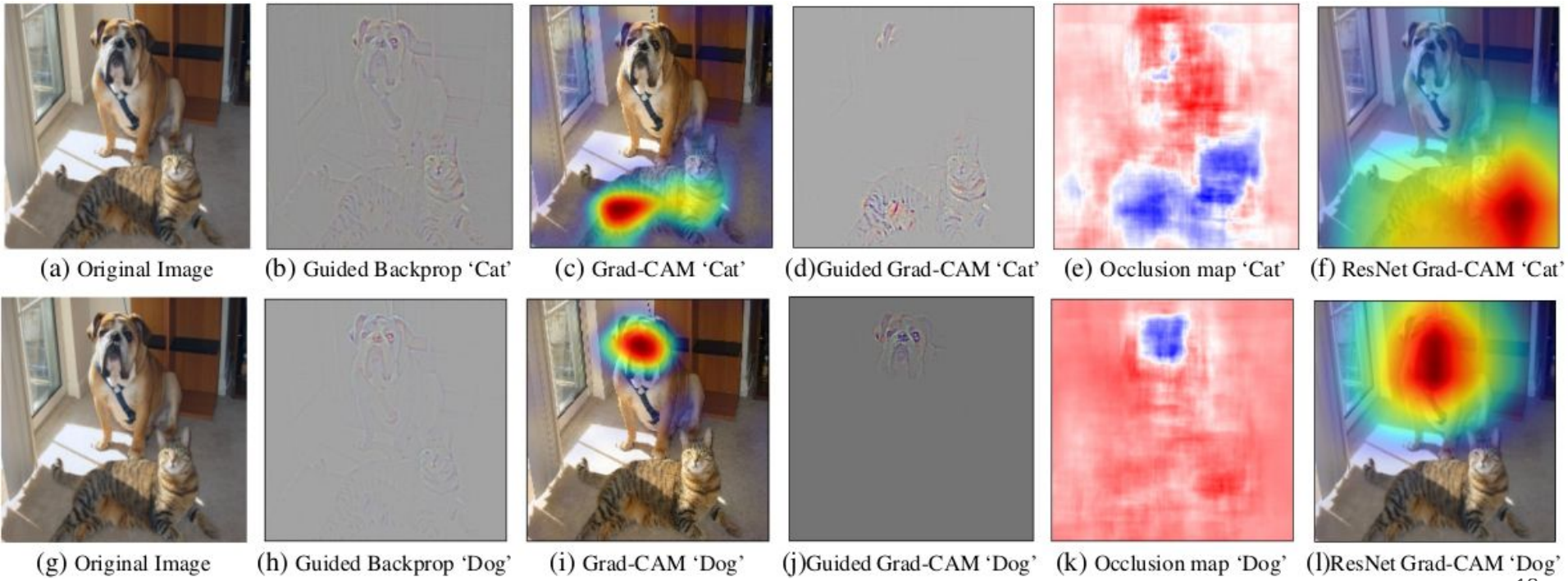
- Must also learn the w's
  - Could formulate this as a multi-objective learning problem
  - But now have a trade-off between prediction accuracy and explainability
- This solution is specific to the classification problem
- Treats positive and negative influences in the same way (just with different signs)

# Grad-CAM

Selvaraju et al., 2019

- Derive the weights from the network directly (no learning)
- Can be applied to networks that solve a range of problems

# Query-Specific Activation



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d)Guided Grad-CAM 'Cat' (e) Occlusion map 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j)Guided Grad-CAM 'Dog' (k) Occlusion map 'Dog' (l)ResNet Grad-CAM 'Dog'

# Counterfactual Classification

Show me the important pixels that are NOT a cat (or dog)

- Use negative weights in the Grad-CAM algorithm

(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

# Semantic Segmentation

- Regions for specific object instances are grown incrementally
- Initial guess at regions are defined by GradCAM activation
- Add neighboring pixels to an existing region if the pixel data are consistent with that of the region

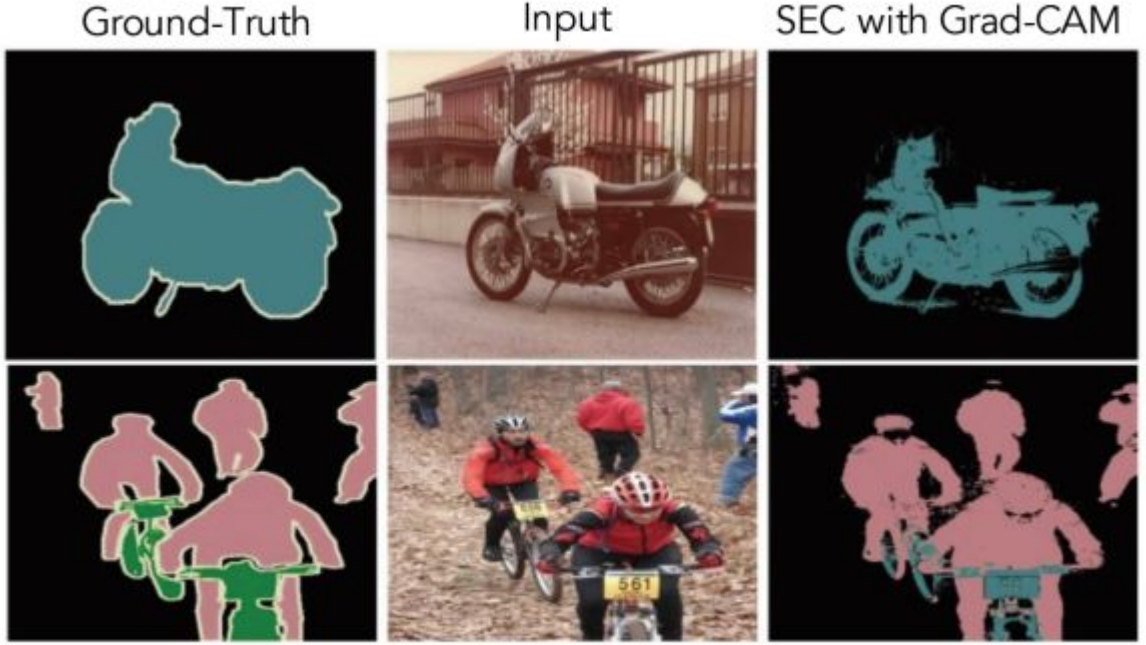
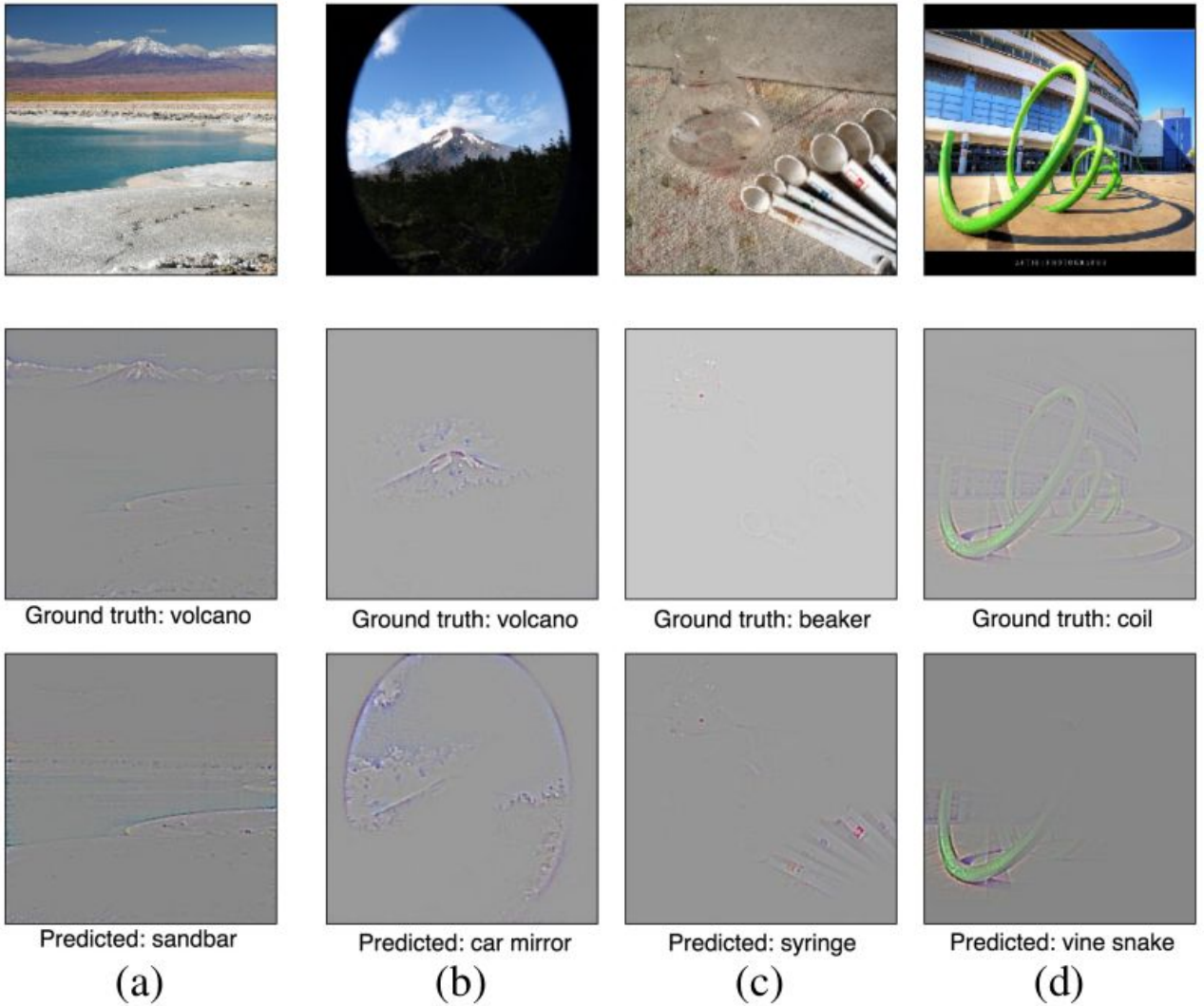



Fig. 4: PASCAL VOC 2012 Segmentation results with Grad-CAM as seed for SEC [32].

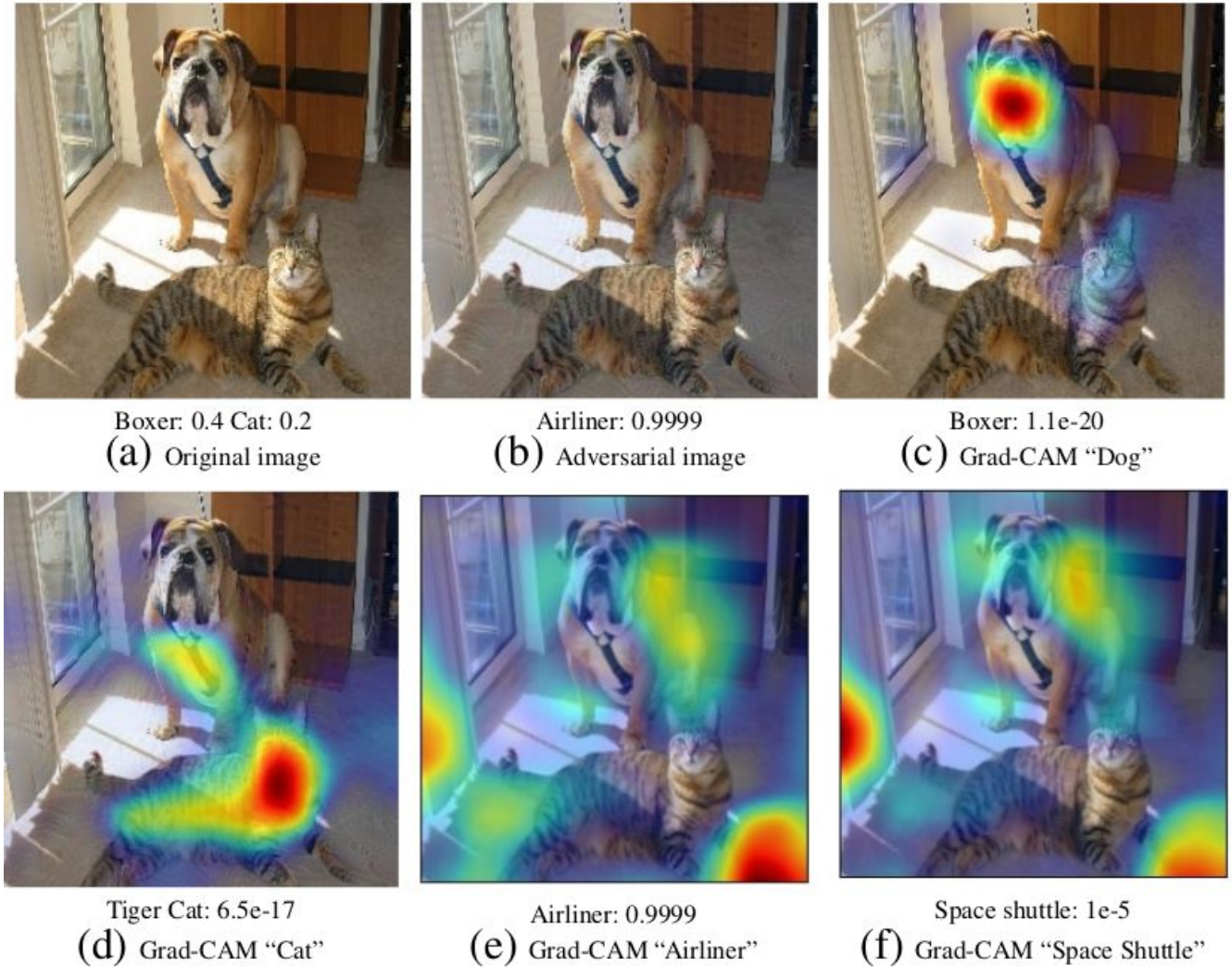
# Analyzing Failure Modes

When a classifier fails, what is it actually paying attention to?



# Adversary Induced Failures

Adversarial attacks: use gradient through the network to alter the pixel values slightly to force a failure



# ImageNet (VGG-16)

How do we tell the difference between doctors and nurses?

Input Image

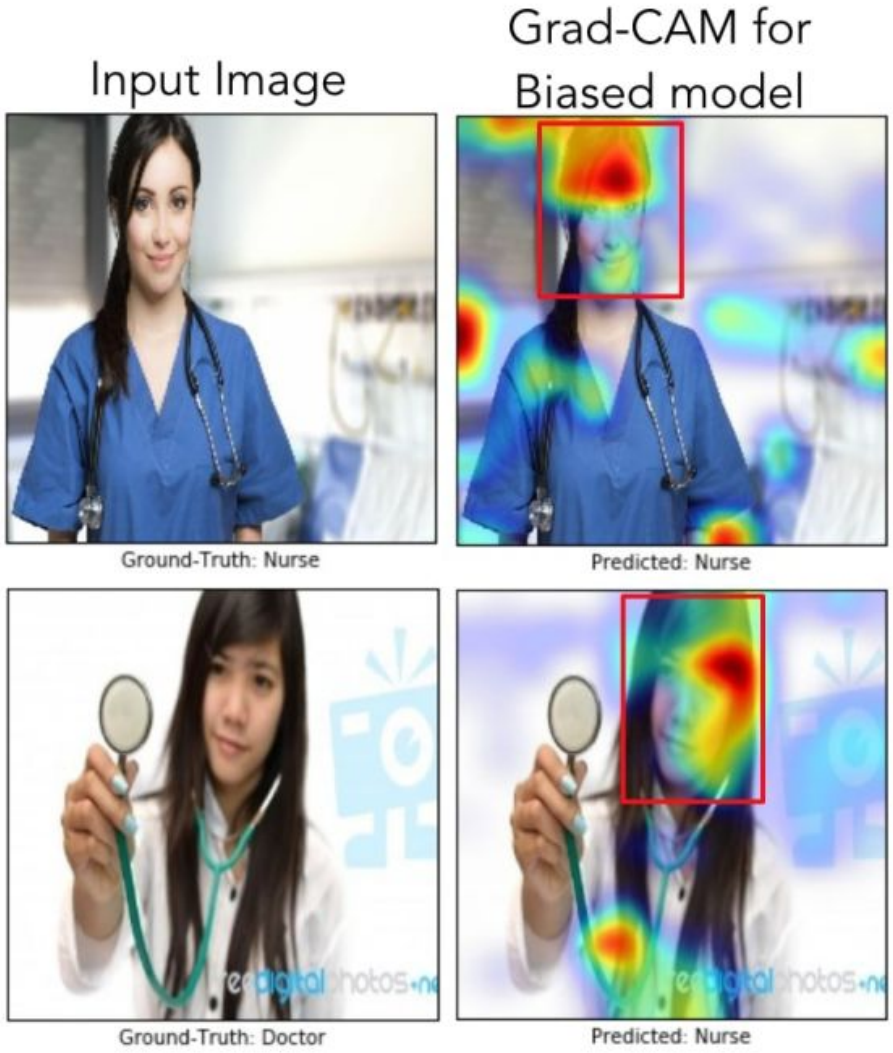
Ground-Truth: Nurse

Ground-Truth: Doctor

# ImageNet (VGG-16)

How do we tell the difference between doctors and nurses?

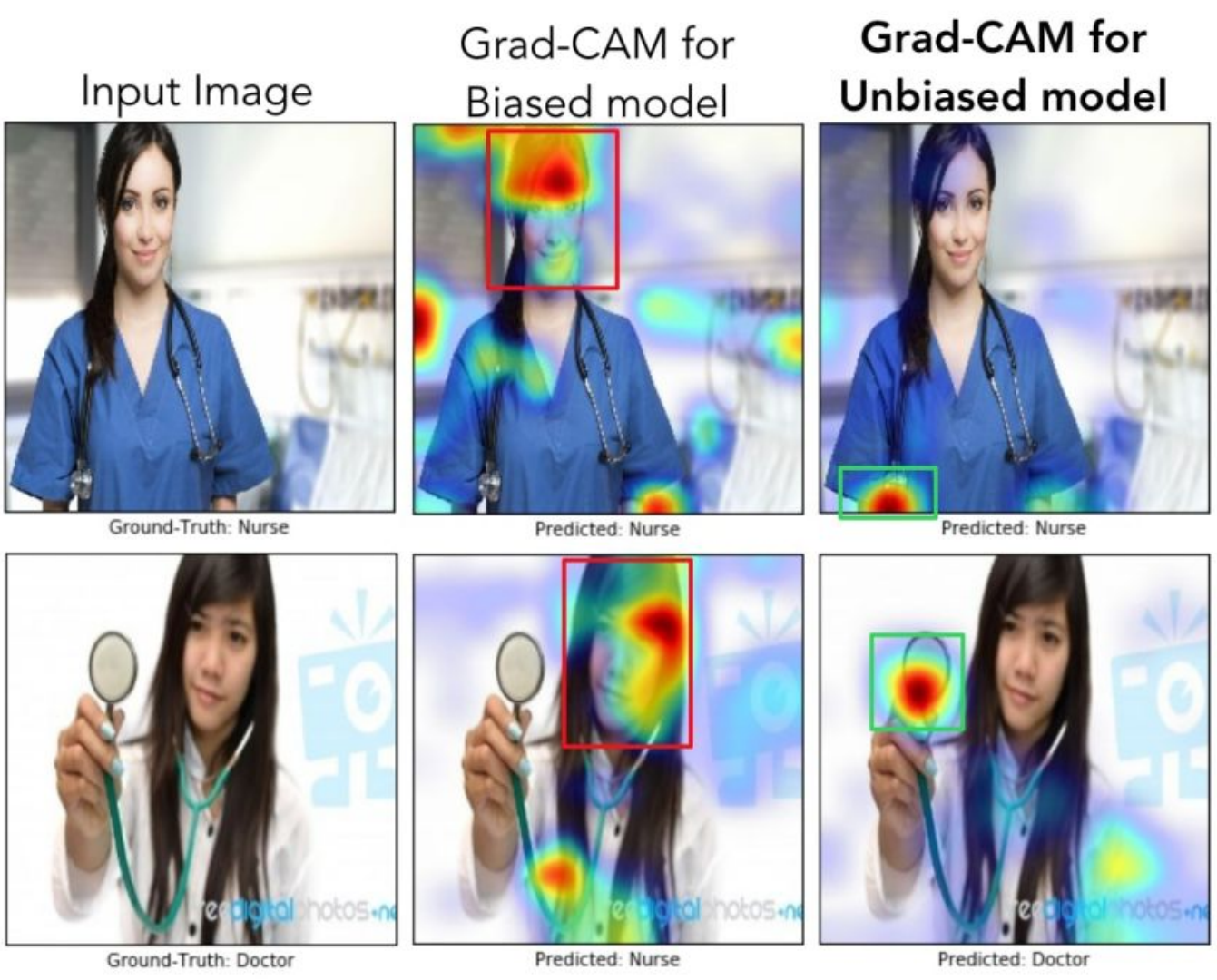
- ImageNet: use face and hair features
- Focused on gender!



# Addressing Bias

Approach:

- Transfer learning: keep the lower layers of VGG-16 and retrain the latter layers
- New training data set: balance gender across doctor and nurse categories


Input Image
Grad-CAM for Biased model
Grad-CAM for Unbiased model
Ground-Truth: Nurse
Predicted: Nurse
Predicted: Nurse
Ground-Truth: Doctor
Predicted: Nurse
Predicted: Doctor

# Grad-CAM

- Fundamental question: what part of the input image is the network “looking at” when solving a prediction problem?
- Apply to any convolutional/pooling layer
- Usable with a wide range of architectures and tasks
- These methods also give the model a way to “point” at specific region of the image that contains the set of pixels involved in the classification decision
    - Could be used for higher-level processing

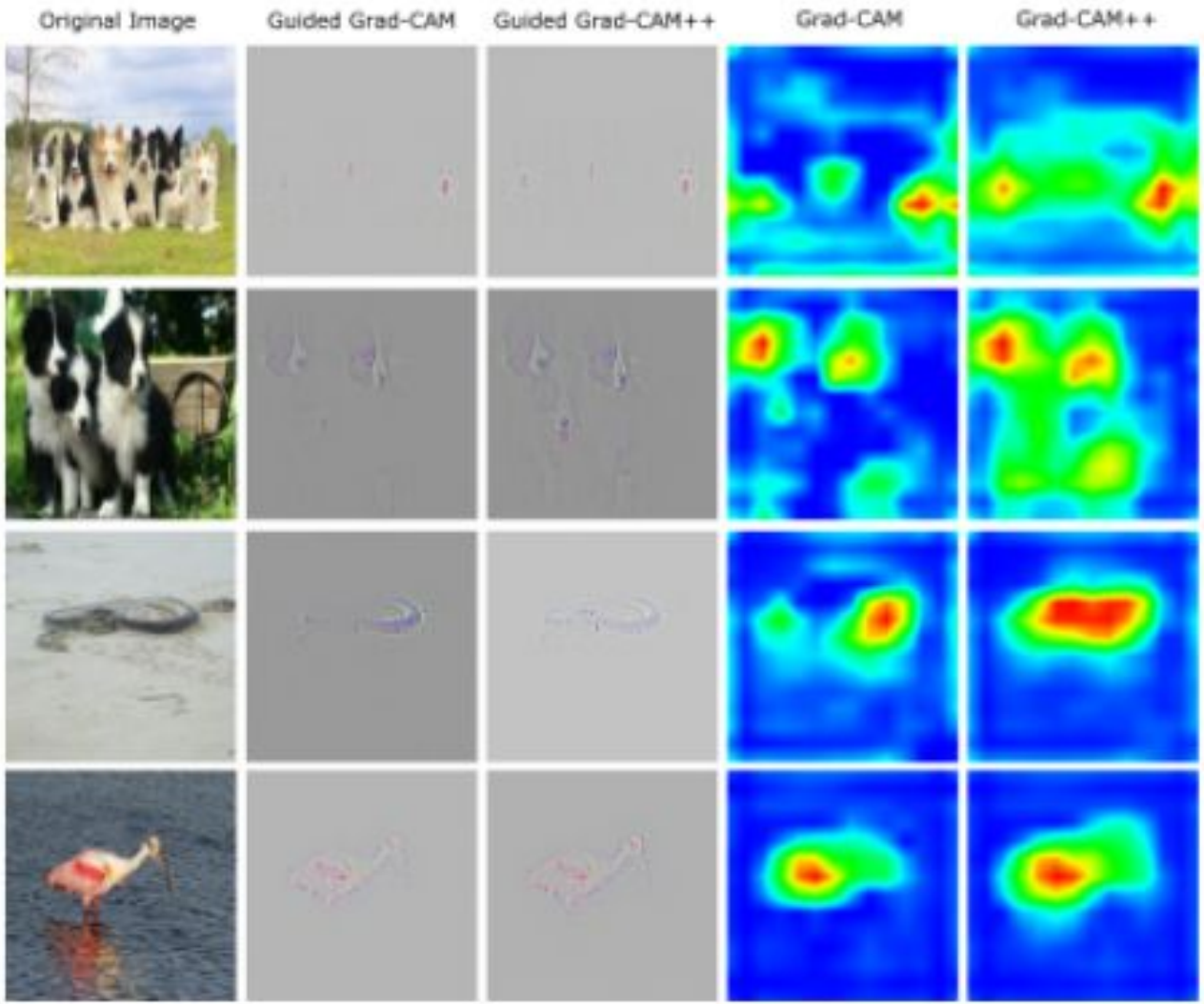
# Grad-CAM Challenges

- For cases with multiple instances of the query object: can give much higher weight to the largest of the instances
- Does not deal with occlusions very well

# Grad-CAM++

- Each pixel has its own weight
- Only focus on positive (or negative) saliencies

Original Image


Original Image
Guided Grad-CAM
Guided Grad-CAM++
Grad-CAM
Grad-CAM++

# Tensorflow Support

- tf-keras-vis: many different methods for visualizing what networks are paying attention to

# References

- A really nice synthesis by V. N. Balasubramanian:
  - https://www.youtube.com/watch?v=VmbBnSv3otc

# Final Thoughts

We want our models to “explain” why they have made specific decisions

- Allows us to bring our own intution/domain knowledge to evaluate what is happening inside of these complex models
- But:
  - Different approaches often give us very different answers
  - There is a lot of **confirmation bias** in how we evaluate these methods