Visualizing Gradient Attribution Methods for Deep Neural Networks

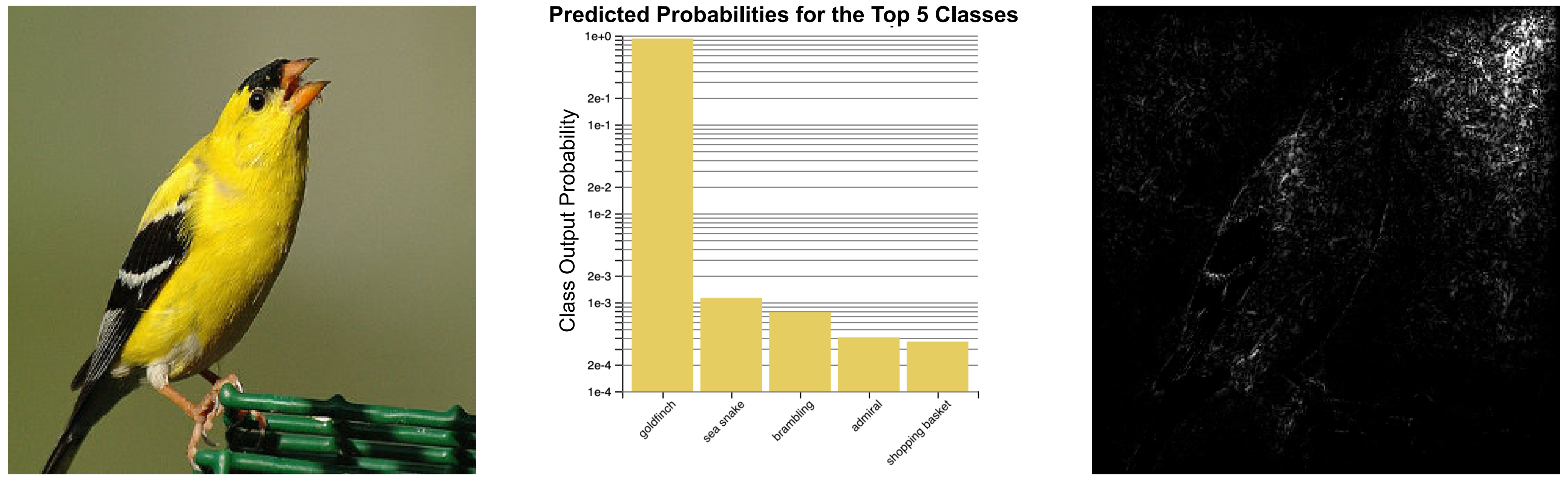
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Pascal Sturmfels

Lee Lab
Explainable AI for Science And Medicine

Deep Neural Networks Achieve High Accuracy - But How?

- State of the art image classification models achieve over 80% accuracy on ImageNet
- But why do networks make the predictions they do? Are they learning meaningful relationships, or learning from background noise?
- Integrated gradients[1] is a method to interpret such networks - it provides a numerical value for each pixel indicating how important that pixel was in making the prediction
- In our example on the left, the network seems to be using the background pixels more than the bird to make the prediction - why?

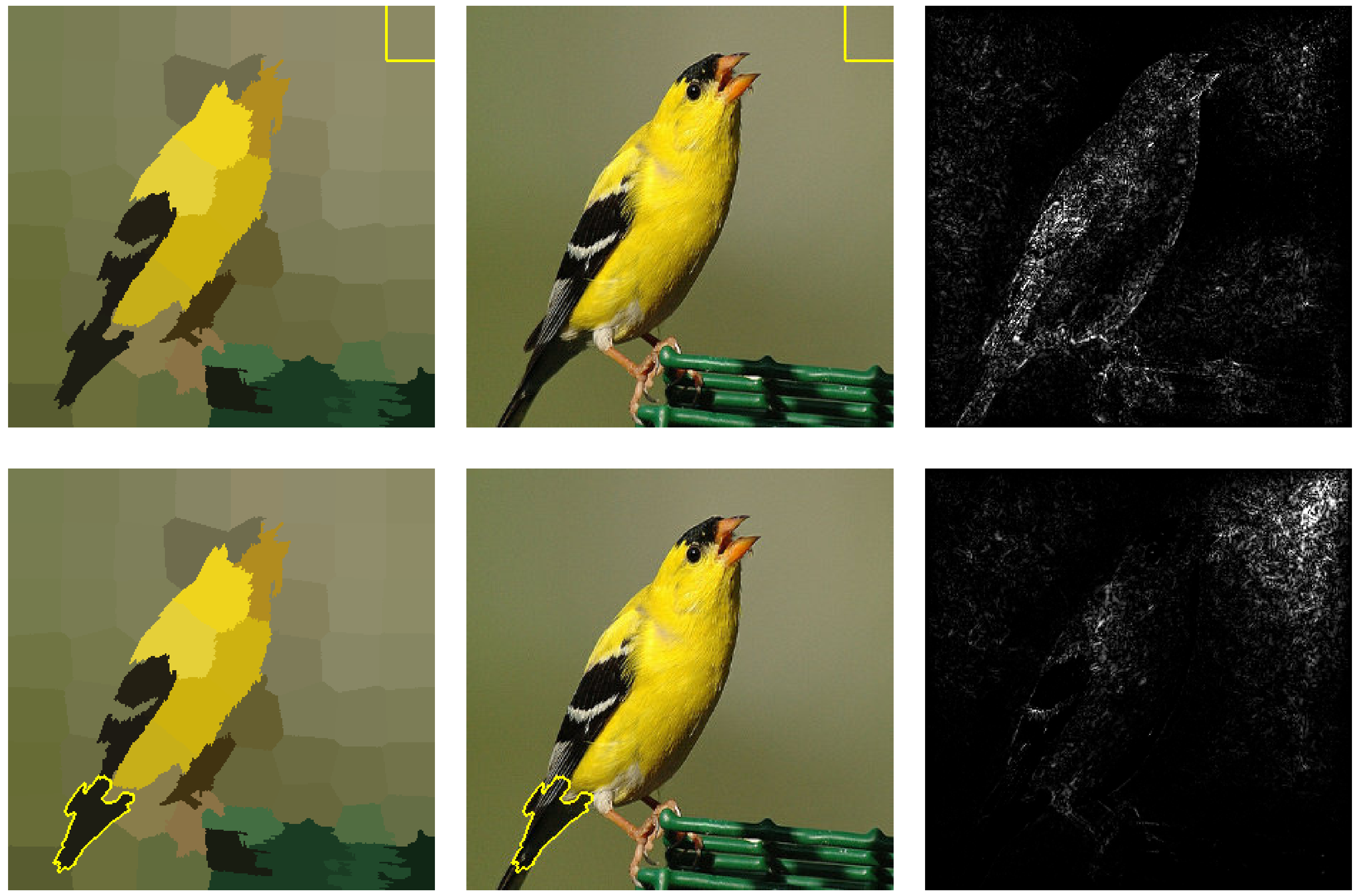


Left: an image from the ImageNet 2012 validation set
Middle: predicted probabilities for the top 5 classes
Right: pixel attribution map generated by integrated gradients

[1] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Discovering Flaws in Network Interpretations using Visualization

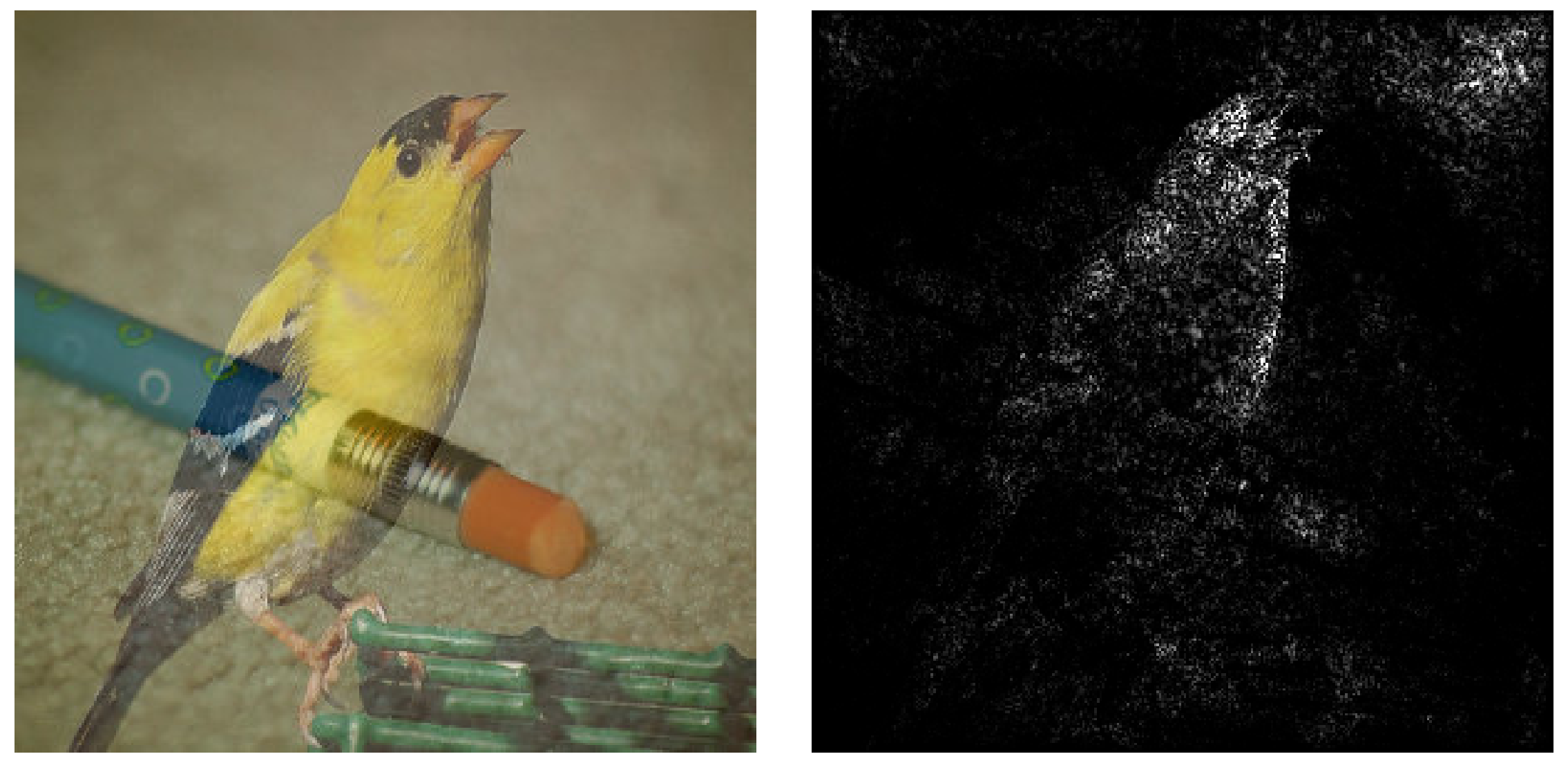
- In order to interpret networks, integrated gradients needs a choice of "reference" that models lack of information
- Depending on the reference, the interpretation can change drastically - we can discover flaws in the predominant method for interpreting deep networks using novel visualizations



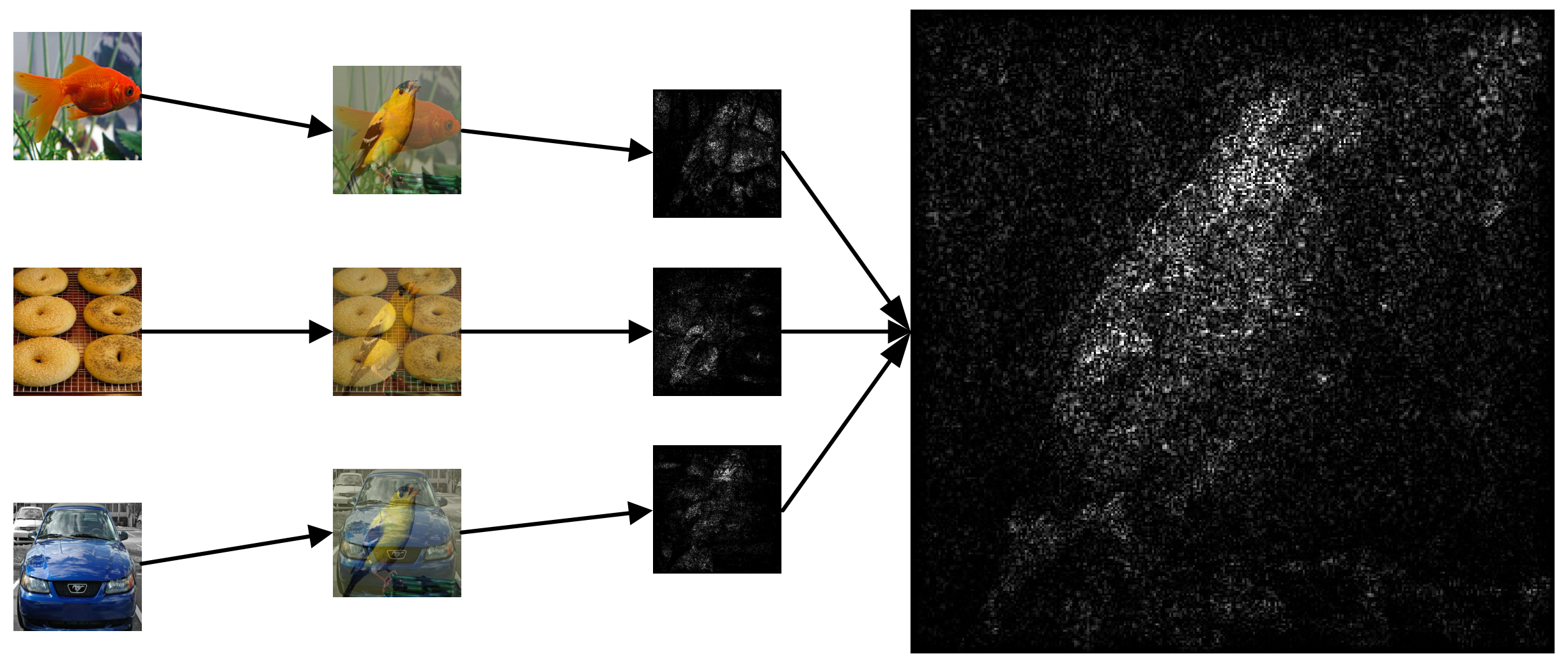
From left to right: the first image represents a segmentation of the test image, along with a highlighted color. That highlighted color is used as a reference for integrated gradients. The second image is the test image with the same region highlighted. The third image is the map of which pixels are important in the image. The two rows represent two different choices of reference. Notice how the attribution map changes depending on the choice of reference color.

Expected Gradients: a New Way to Interpret Neural Networks

- Instead of using a constant color as a reference, we can use a different image. This asks a counter-factual question: "Why do you see a bird and not a pencil?"
- We can extend the idea by using multiple images in the training dataset, which inspires our method: expected gradients.
- Expected gradients gets which pixels are important relative to the training distribution - visualizations demonstrate it better accumulates gradients important to the network



Integrating between a pencil and a goldfinch. The pixel attributions highlight both pixels in the bird and pixels in the pencil.



Expected gradients, visualized. First, we draw samples from the training distribution. Then, we interpolate between those samples and the original image. We next calculate the gradients at the interpolated images, and finally we take the mean over all of them in order to get the expected gradients attribution map.

Acknowledgements

The method is part of a larger work, in large part aided by Scott Lundberg, who also deserves thanks for ideas and advice regarding this visualization. Thanks as well to Joe Janizek and Gabe Erion, who were responsible for much of the conference paper we introduced this method in. Similarly, thanks to Su In Lee for feedback regarding the paper.