Visualizing Gradient Attribution Methods for Deep Neural Networks

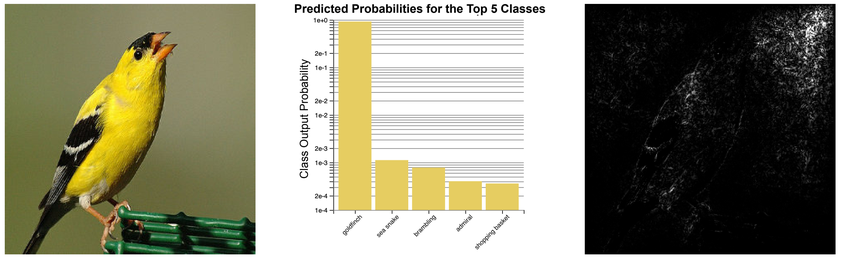
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Pascal Sturmfels

Lee Lab
Explainable AI for Science And Medicine

Deep Neural Networks Achieve High Accuracy - But How?

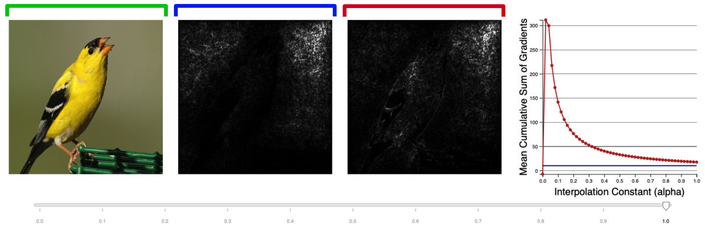
- State of the art image classification models achieve over 80% accuracy on ImageNet
- But why do networks make the predictions they do? Are they learning meaningful relationships, or learning from background noise?



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

- Integrated gradients[1] is a method to interpret such networks - for each pixel, gives a value indicating how influential that pixel was towards making the output prediction
- Interactive visualization of the underlying equation helps us better understand the method - and its flaws

$$\text{IntegratedGradients}_i(x, x') = \int_{\alpha=0}^{1} (x_i - x'_i) \times \frac{\delta f(x' + \alpha \times (x - x'))}{\delta x_i} \delta\alpha$$

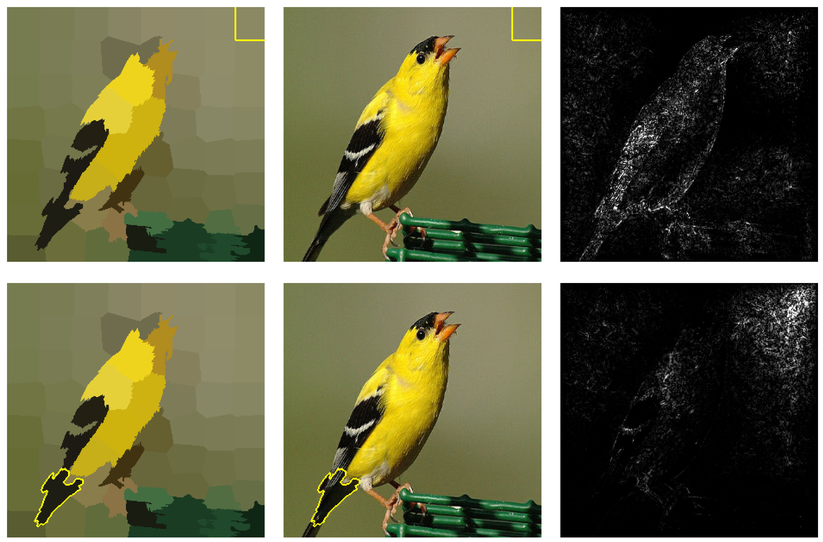


The integrated gradients equation visualized. Each picture corresponds to the part of the equation indicated by the colored brackets.

[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" x' that models lack of information
- The color black is often chosen - but black is a meaningful color.
- Depending on the color of reference, the interpretation can change drastically - we can discover flaws in the predominant method for interpreting deep networks using interactive visualizations



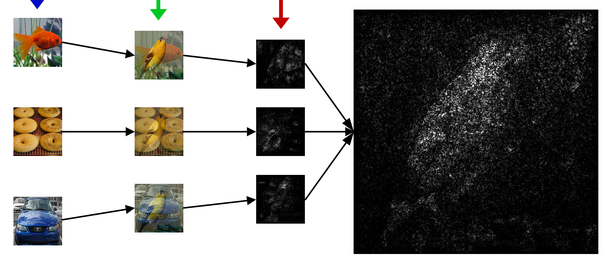
From left to right: the first image represents a segmentation of the test image, along with a highlighted color. That highlighted color is used as a reference for integrated gradients. The second image is the test image with the same region highlighted. The third image is the map of which pixels are 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 Novel Attribution Method

- 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

$$\text{ExpectedGradients}_i(x) = \underset{x' \sim D, \alpha \sim U(0,1)}{E} \left[(x_i - x'_i) \times \frac{\delta f(x' + \alpha \times (x - x'))}{\delta x_i} \right]$$

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

My lab colleagues and I introduced expected gradients in a recent conference submission. Thanks to my collaborators Scott Lundberg, Joe Janizek, Gabe Erion and Su In Lee. Additional thanks to Scott for ideas regarding the visualizations presented here.