Google Research

# Guided Integrated Gradients: an Adaptive Path Method for Removing Noise

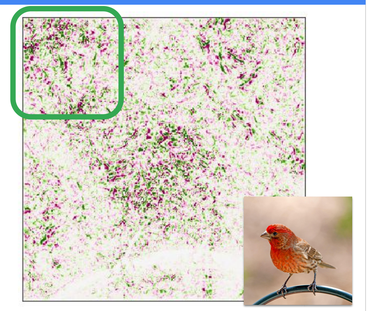
Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, Tolga Bolukbasi
Google Research

CVPR VIRTUAL JUNE 19-25

## Motivation

**Attribution methods** link a deep neural network's prediction to the input features that most influence that prediction.

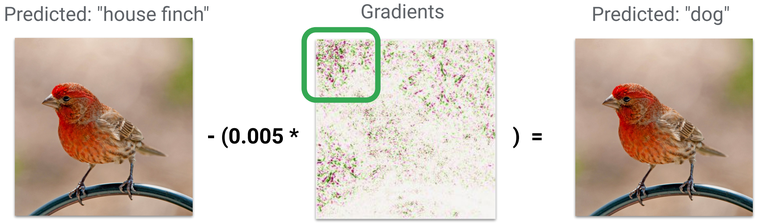
**Integrated Gradients (IG)** method is a well established axiomatic attribution method with a solid theoretical foundation but it often produces noisy attributions.



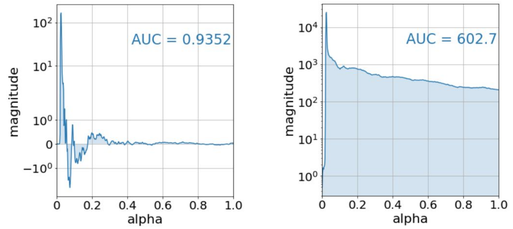
**Left:** Integrated Gradients attribution for the model prediction "house finch".
**Right:** The input image.

## Noise Accumulation

IG attribution is significantly impacted by high gradients along the integration path, even when these gradients are irrelevant to the final model prediction.




The average magnitude of gradients **(below, right)** is much higher than the magnitude of directional derivative **(below, left)** that actually change the prediction of the model.




## Adaptive Path Methods

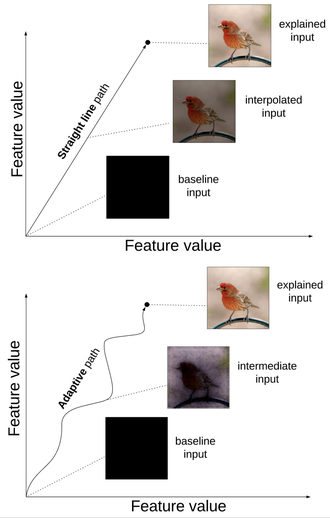
**Fixed path methods** such as Integrated Gradients construct a fixed integration path using the input image only.

$$IG_i(x) = \int_{\alpha=0}^{1} \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} d\alpha$$

**Adaptive path methods** are a new family of attribution methods that can take advantage of the knowledge of the model to dynamically construct a path that meets a desirable objective.

**Definition:**

$$a_i^{\gamma^F}(X^I) = \int_{\alpha=0}^{1} \frac{\partial F(\gamma^F(\alpha))}{\partial \gamma_i^F(\alpha)} \frac{\partial \gamma_i^F(\alpha)}{\partial \alpha} d\alpha$$




## Guided Integrated Gradients

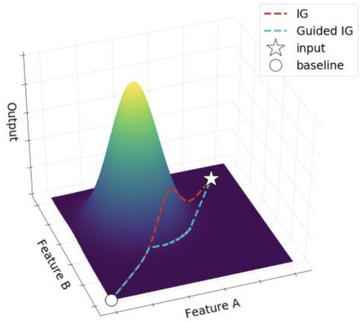
**Objective:** Find a path that minimizes the effect of irrelevant high gradients on result attribution.

**Algorithm:**

- Start from the baseline and move toward the input (same as IG).
- Move only in the direction of features that have the lowest absolute value of associated partial derivatives and are not equal to the input.

**Properties:**

- Satisfies a set of axioms.
- Has the same asymptotic time complexity as IG.



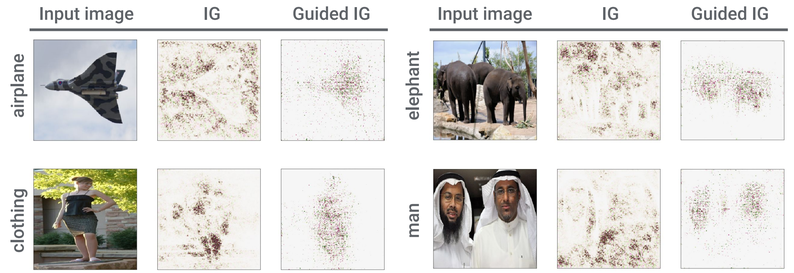

Schematic comparing IG and Guided IG. By moving in the direction of lowest associated partial derivatives, Guided IG minimizes the influence of high gradient examples.

## Results

Across 5 models and 3 datasets (natural images and diabetic retinopathy), Guided IG outperforms other pixel-based attribution methods, and improves other methods that are use pixel-based attribution maps as input **(right).**

We confirm these quantitative findings by observing Guided IG reducing noise in images compared to Integrated Gradients (IG) **(below).**

| (AUC) | ImageNet | | | Open Images | DR |
|---|---|---|---|---|---|
| Method | MobileNet | Inception | ResNet | ResNet | Inception |
| Edge | 0.611 | 0.610 | 0.611 | 0.606 | 0.643 |
| Gradients | 0.614 | 0.634 | 0.650 | 0.505 | 0.801 |
| IG | 0.629 | 0.655 | 0.669 | 0.557 | 0.833 |
| Blur IG | 0.652 | 0.662 | 0.663 | 0.619 | 0.830 |
| GIG(0) | **0.705** | **0.712** | **0.711** | **0.630** | 0.619 |
| GIG(20) | 0.691 | 0.696 | 0.706 | 0.624 | **0.863*** |
| GradCAM | 0.776 | 0.761 | 0.755 | 0.474 | 0.837 |
| Smoothgrad | | | | | |
| +IG | 0.742 | 0.773 | 0.781 | 0.662 | 0.637 |
| +GIG(0) | 0.745 | 0.776 | 0.776 | 0.649 | 0.632 |
| +GIG(20) | **0.767** | **0.795** | **0.799** | **0.685** | **0.645** |
| XRAI | | | | | |
| +IG | 0.731 | 0.765 | 0.762 | 0.631 | 0.793 |
| +GIG(0) | **0.838*** | **0.829*** | **0.821*** | 0.718 | 0.630 |
| +GIG(20) | 0.808 | 0.819 | 0.809 | **0.719*** | **0.831** |




## Summary

- **Adaptive path methods** are attribution methods that are conditioned on input and model.
- **Guided IG** is a new axiomatic attribution method that minimizes effect of undesirable (including adversarial) gradients.
- The method produces better scores than other path methods according to weak localization and PIC metrics.

https://github.com/PAIR-code/saliency