Sanity Checks for Saliency Maps

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim
AI models can be ‘black boxes’ and saliency maps try to highlight which input features (e.g. pixels) matter the most for a given prediction.

Examples include Gradients, Grad-CAM, Integrated Gradients, Guided Backprop…

Relying solely on visual appeal (the ‘map’) can be misleading.
Some saliency methods (e.g. Guided Backprop) look very similar to classical edge detectors.

Edge detectors require no training data or labels.

Visual similarity could be misleading if map is just highlighting edges.
Background - Saliency Maps vs. Edge Detection

| Original Image | Gradient | SmoothGrad | Guided BackProp | Guided GradCAM | Integrated Gradients | Integrated Gradients SmoothGrad | Gradient Input | Edge Detector |
|----------------|----------|------------|-----------------|---------------|---------------------|-----------------------------|----------------|---------------|
| Junco Bird     |          |            |                 |               |                     |                             |                |               |
|                |          |            |                 |               |                     |                             |                |               |
| Corn           |          |            |                 |               |                     |                             |                |               |
|                |          |            |                 |               |                     |                             |                |               |
| Wheaten Terrier|          |            |                 |               |                     |                             |                |               |
|                |          |            |                 |               |                     |                             |                |               |
Do saliency methods reflect model-data relationships, or do they just highlight superficial cues (like edges)?
Approach

Model Parameter Randomisation Test

Data Randomisation Test

Gradient, SmoothGrad, Guided BackProp, Guided GradCAM, Integrated Gradients, IGSG, Gradient⊙Input

Inception v3 (ImageNet), CNNs on MNIST/Fashion-MNIST, MLP

Visual inspection, Spearman rank correlation (with/without absolute values), Structural Similarity Index (SSIM) and Histogram of Gradients (HOG) similarity
Model Parameter Randomisation Tests

- Randomise model weights (top layer → bottom layer)
- Cascading vs. Independent
- Generate saliency maps after each randomisation step
Model Parameter Randomisation Tests - Cascading
Model Parameter Randomisation Tests - Cascading
Model Parameter Randomisation Tests - Independent
Data Randomisation Test

1. Shuffle training labels
2. Train a new model to fit random labels
3. Compare saliency maps from correctly-labelled model to randomly-labelled model
Data Randomisation Test
Key Findings

Saliency methods differ in sensitivity, some strongly reflect the learned parameters and data labels while others appear nearly unchanged when the model or labels are randomised.

Visual similarity ≠ True explanation

Simple checks (randomisation tests) can reveal if a method genuinely depends on training.
Key Findings

‘Architecture as a Prior’ – design of neural network can embed biases about how data should be processed

Element-wise input ⊙ gradient (or similar approaches) can display the input’s outline even if gradient is random
Related Work

| Name, Description and Main Explanation Types | References |
|---------------------------------------------|------------|
| **Related Work**                            |            |
| **Consistency (Section 6.5)**               |            |
| Implementation Invariance – Feature Importance | [1, 2, 4]  |
| Evaluate whether the explanation method is invariant to specific implementations of the predictive model by validating whether two implementations that give the same output for an input, also get the same explanation. |            |
### Related Work

#### Table 3. Continued

| Feature Importance, Heatmap, Graph, Text, Localization, Decision Rules, White-box model |
|------------------------------------------|
| Stability for Slight Variations          |
| **Measure the similarity between explanations for two slightly different samples.** |
| Small variations in the input, for which the model response is nearly identical, should not lead to large changes in the explanation. |
| References: [10, 15, 16, 18, 20, 21, 22, 23, 24, 25, 26, 27, 31] |
| Fidelity for Slight Variations – Decision Rules, White-box model |
| **Measure the agreement between interpretable predictions for original and slightly different samples:** an explanation for original input $x$ should accurately predict the model’s output for a slightly different sample $x'$. |
| References: [14, 20] |
| Connectedness – Prototypes, Representation Synthesis |
| **Measure how connected a counterfactual explanation is to samples in the training data. Ideally, the counterfactual is not an outlier, and there is a continuous path between a generalized counterfactual and a training sample.** |
| References: [12, 15, 20] |
| **CONTINUITY (Section 6.1)** |
| **Target Sensitivity – Heatmap** |
| The explanation for a particular target or model output (e.g. class) should be different from an explanation for another target. |
| References: [18, 20, 25, 26, 27, 31] |
| **Target Discriminativeness – Disentanglement, Representation Synthesis, Text** |
| The explanation should be target-discriminative such that another model can predict the right target (e.g. class label) from the explanation, in either a supervised or unsupervised fashion. |
| References: [15, 25, 26, 27, 28, 31, 33] |
| **Randomization Check – Feature importance, Heatmap, Localization** |
| Randomly change labels in a copy of the training dataset, train a model on this randomized dataset and check that the explanations for this model on a test set are different from the explanations for the model trained on the original training data. |
| References: [1, 13, 14, 124] |
| **CONTRASTIVITY (Section 6.5)** |
| **Covariate Homogeneity** |
| Prototypes, Disentanglement, Localization, Heatmap, Representation Synthesis |
| Evaluate how consistently a covariate (i.e. feature) in an explanation represents a predefined human-interpretable concept. |
| References: [18, 26, 29, 30, 31, 32, 33] |
| **Covariate Regularity – Decision Rules, Feature Importance** |
| Evaluate the regularity of an explanation by measuring its Shannon entropy, in order to quantify how noisy the explanation is and how easy it is to memorize the explanation. |
| References: [20, 304] |
| **COVARIATE COMPLEXITY (Section 6.6)** |
| **Size** |
| Feature importance, Heatmap, Decision Rules, Prototypes, Text, Graph, Localization, White-box model, Representation Synthesis |
| **Total size (absolute) or sparsity (relative) of the explanation.** |
| References: [1, 15, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33] |
| **Redundancy – Feature importance, Decision Rules, Text, White-box model** |
| Calculate the redundancy or overlap between parts of the explanation. |
| References: [140, 151, 154] |
| **Counterfactual Compactness – Prototypes, Representation Synthesis, Text** |
| Given a counterfactual explanation showing what needs to be changed in the input in order to change the prediction of the predictive model, measure how much needs to be changed. |
| References: [1, 25, 151, 154] |
Positives

Highly quantitative

Seminal

Easy to replicate
Negatives

Focus only on images

Not many architectures tested
Future Work

Apply tests to other modalities

Could combine with ablation or concept-based approaches to investigate causality

Test how saliency changes under partial label noise