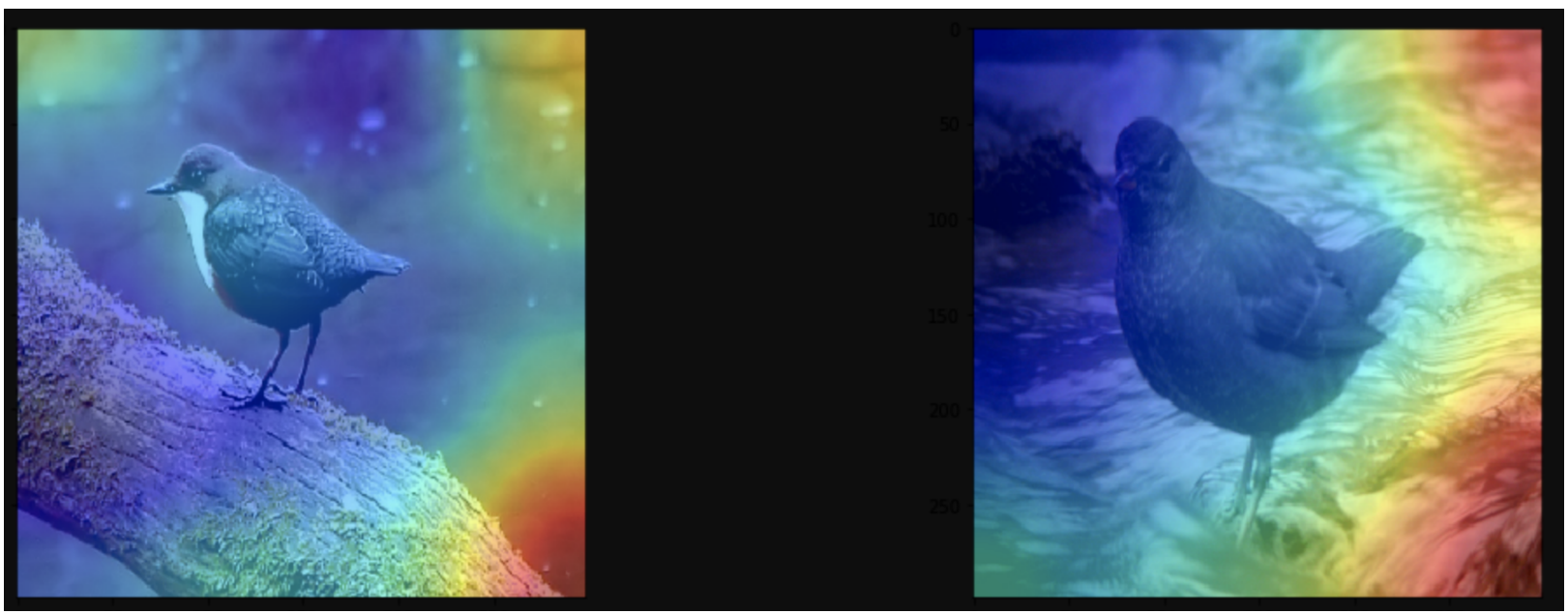
Mini-Survey of Attribution



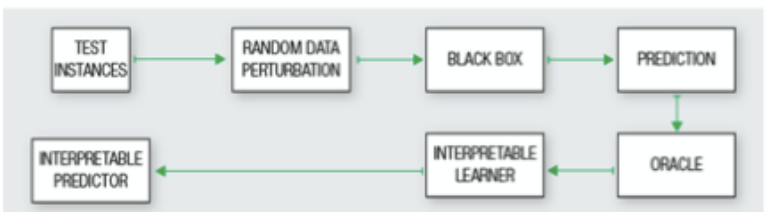
Why?

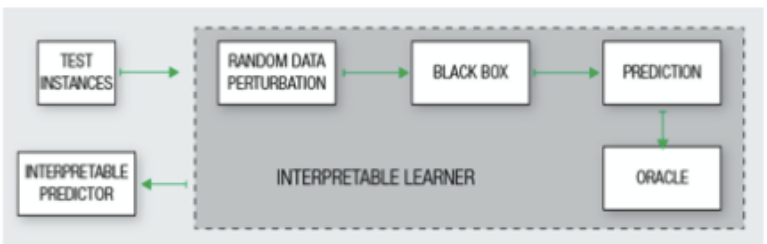
- Interpretability of Deep Learning
- Explain what input explains decisions

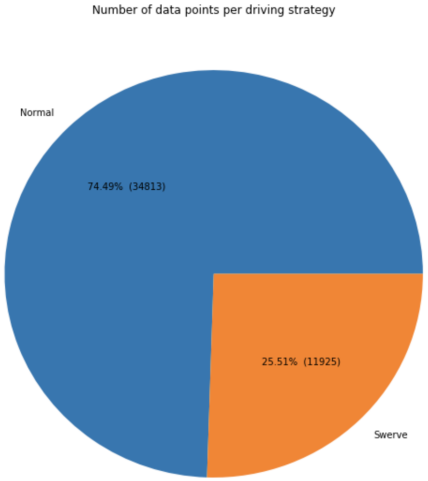

Background

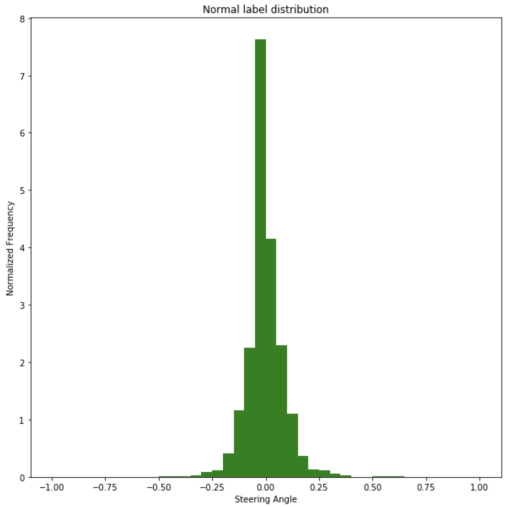
Some Types of Attribution

- Data Analysis
- Relationship between inputs and outputs
  - Pertinent parts of the dataset
  - Dataset relationship to the output
  - Relationship between model layers and the output
- Feature Analysis

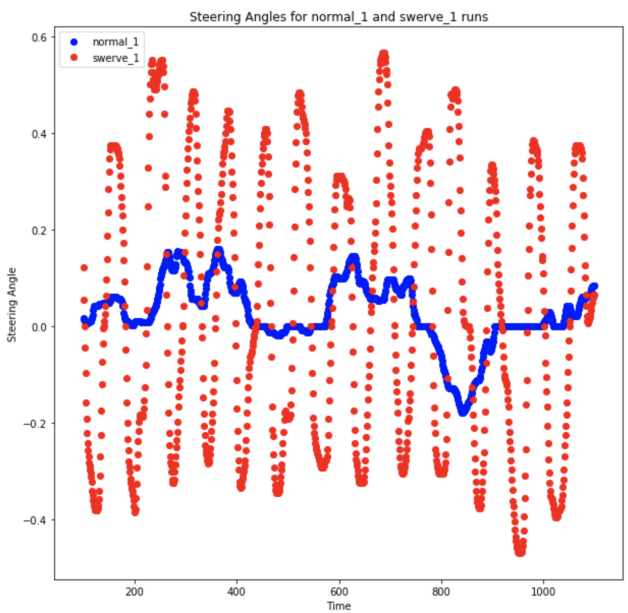

TEST INSTANCES
RANDOM DATA PERTURBATION
BLACK BOX
PREDICTION
INTERPRETABLE PREDICTOR
INTERPRETABLE LEARNER
ORACLE

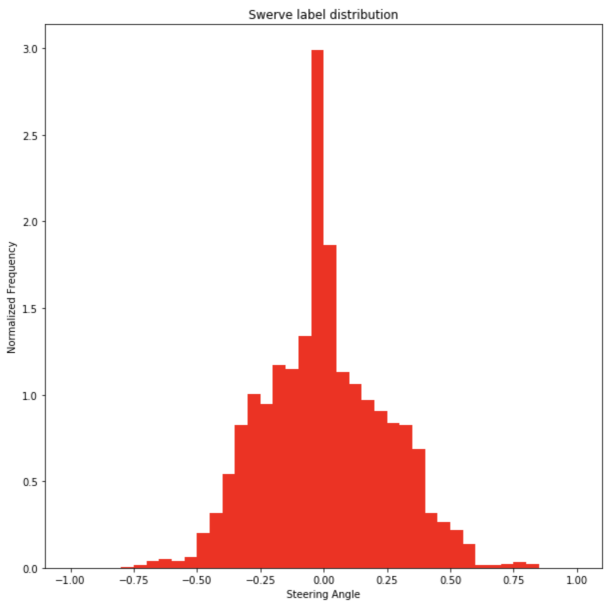

TEST INSTANCES
RANDOM DATA PERTURBATION
BLACK BOX
PREDICTION
INTERPRETABLE PREDICTOR
INTERPRETABLE LEARNER
ORACLE





Data Analysis





Bias?

Distribution

Attribution Methods

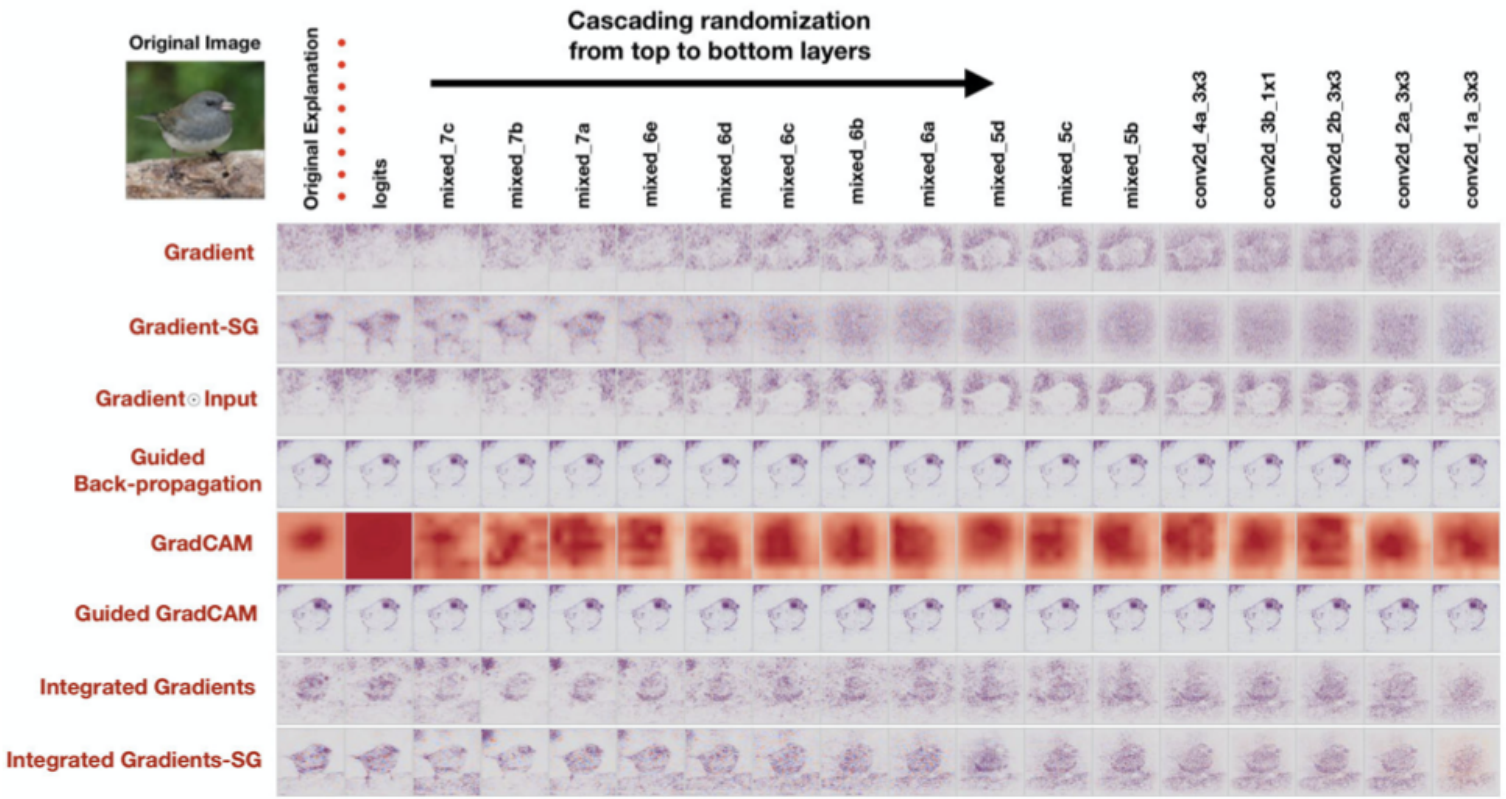
- Model relationship tests
- Perturbation Methods
- Gradient methods
- Other methods

Model relationship tests

- Model Parameter Randomization Test
  - Cascading Randomization
  - Independant Randomization
- Data Randomization Test

Cascading Randomization

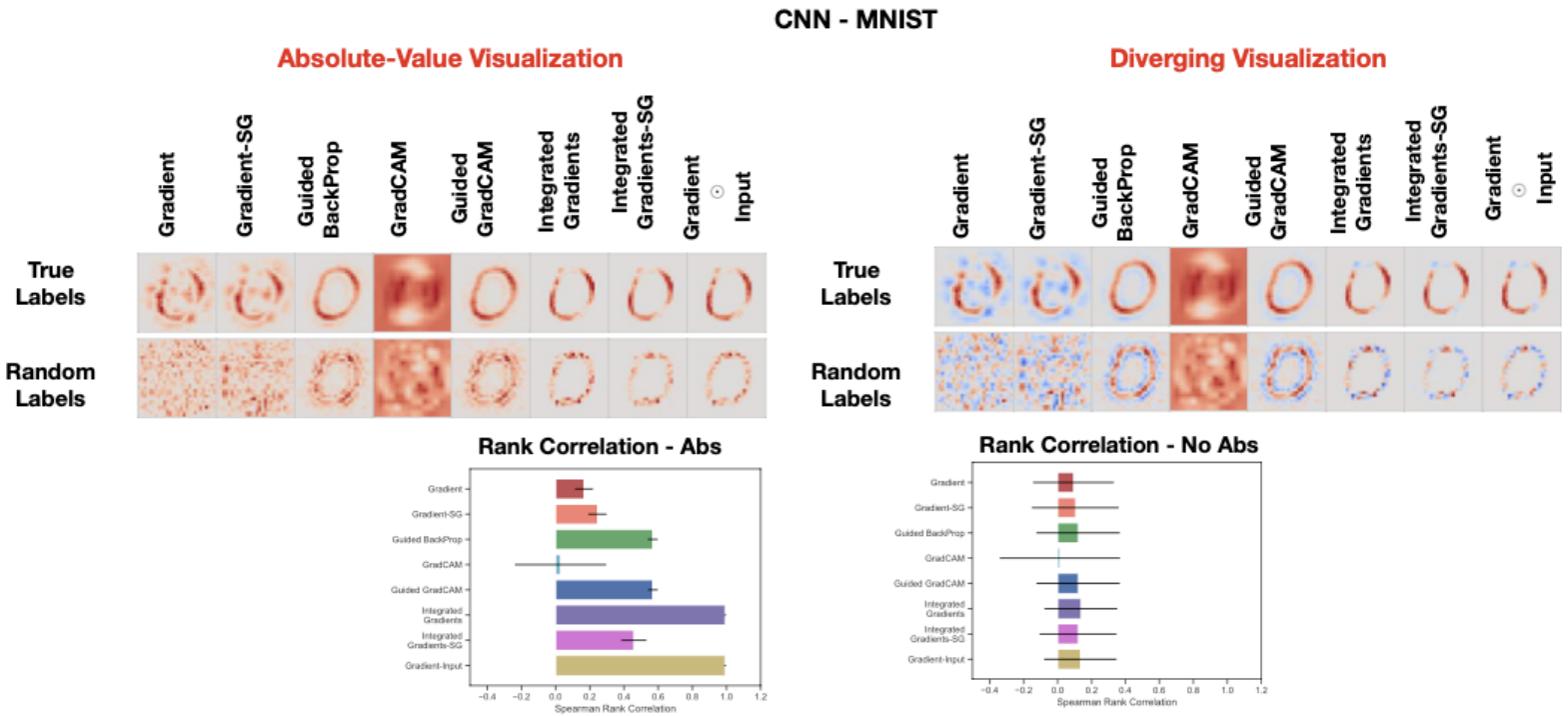
- The weights of a model are randomized over time starting from the top layers and moving down to the bottom ones.
- This test shows the sensitivity of an attribution method to the model's parameters.


Original Image
Cascading randomization
from top to bottom layers
Original Explanation
logits
mixed_7c
mixed_7b
mixed_7a
mixed_6e
mixed_6d
mixed_6c
mixed_6b
mixed_6a
mixed_5d
mixed_5c
mixed_5b
conv2d_4a_3x3
conv2d_3b_1x1
conv2d_2b_3x3
conv2d_2a_3x3
conv2d_1a_3x3
Gradient
Gradient-SG
Gradient ⊙ Input
Guided
Back-propagation
GradCAM
Guided GradCAM
Integrated Gradients
Integrated Gradients-SG

Independant Randomization

- This is done by performing randomization layer-by-layer rather than by weight

- This gives a more granular indication of dependency for an attribution method by the order of the layer.



Data randomization results

Perturbation

- Type of function which compares two networks
  - The original network
  - A network trained on the dataset where relevant features have been altered
    - Masked
    - Obscured
    - Removed
    - Biased

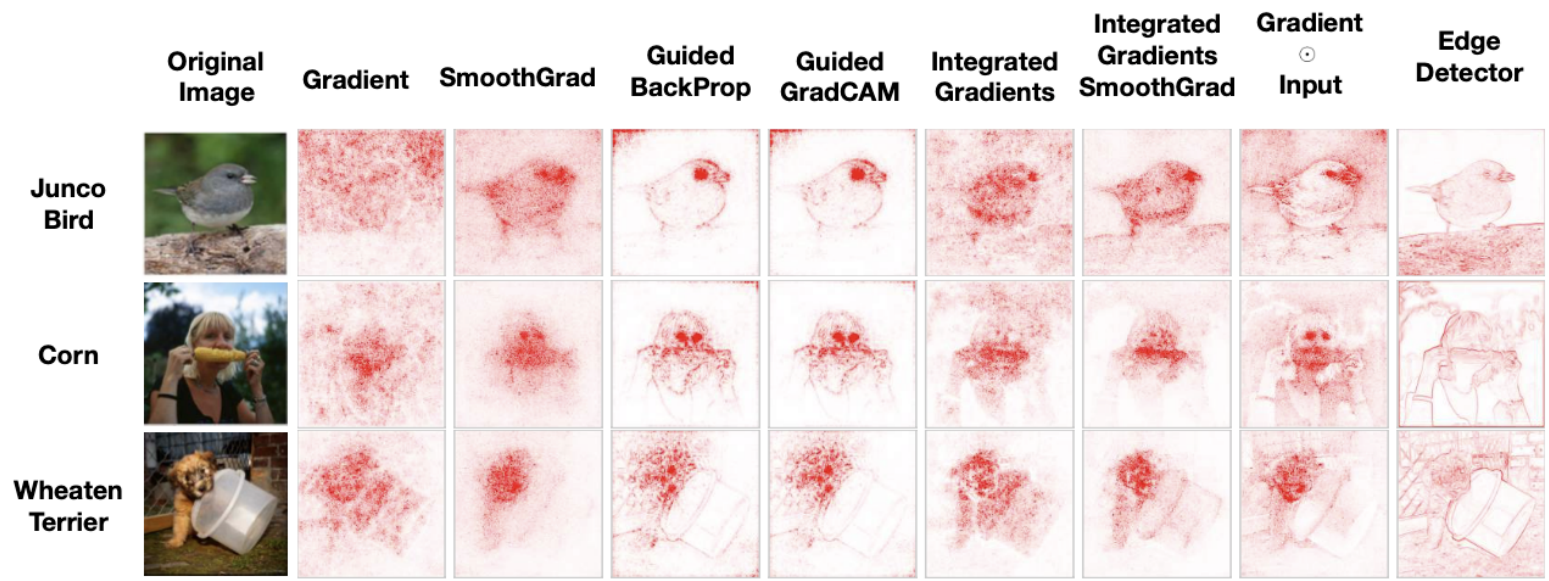
Gradient Methods

- Make use of the gradient based methods in the back-propagation step to attempt to extract spatial information captured in the network

Saliency Maps

- Compute the absolute partial derivative of the output neuron with respect to the input features

$$\frac{\partial \text{output}}{\partial input}$$


Original Image
Gradient
SmoothGrad
Guided BackProp
Guided GradCAM
Integrated Gradients
Integrated Gradients SmoothGrad
Gradient ⊙ Input
Edge Detector
Junco Bird
Corn
Wheaten Terrier

Issues with Gradient Based Methods

- Many methods partially reconstruct the input data
- Brittle to noise and interference (misleading results)
- Many of the advanced guided methods dont have an adequate relationship between the input data and output nodes of a network
- Some methods (like some saliency maps) may not work with features that have a negative effect on the output

Other Methods

- Tree approximations
- Uncertainty of the models (Sensitivity Analysis)
- Prototype Selection

Issues with Attribution

- The outputs of networks may have dimensions that are much smaller than the inputs
- Noise becomes a factor which may cause artifacting
- The boundaries of visualisation techniques may become distorted due to network shape and feature extractor size
- The final step of attribution may introduce its own bias i.e. upscaling, resizing etc
- Computational cost

Evaluating Attribution

- Take the result of an attribution method and find reasoning from it

Evaluating Attribution

- Generate visualisations which are then interpreted by a human
- Perturbation techniques on top of other techniques to find localised importance from the dataset
- Model performance metric and perturbation
- Calculating similarity between a truth dataset feature map and an extracted feature map (Something like KL-divergence)

KL-Divergence

Kullback and Leibler (1951) (and Belov and Armstrong (2011)) define the KL Divergence as:

$$D(g\|b) = \int_{-\infty}^{+\infty} g(x) \ln \frac{g(x)}{b(x)} \mathrm{d}x$$

where $g(x)$ and $h(x)$ are probability density functions.