Sanity Checks for Saliency Maps

21 Jan 2020

Presenter: Sanchit Sinha

https://qdata.github.io/deep2Read/

Motivation

- **Saliency maps** - to study contribution of various neurons in final outputs/predictions of DNNs
- **Saliency methods** - used for explaining predictions made by neural nets. (Generate Saliency maps)
- Several saliency methods proposed - Guided Backpropagation, Guided GradCAM, Integrated Gradients
- The visual maps generated by the saliency methods tend to be evaluated visually - which is **not completely correct**
- Comparisons to Edge maps are mere **coincidence** in most cases and cannot be used to justify the
- **Which SM to use? Which is the most explainable/robust?**
- Devising experiments on parameters and data to check which saliency method gives the best explanation

Background

- **Need for Saliency methods** - There is no consensus if the models produce explainable predictions
- **Why Explainability?** - Required for debugging, remove bias, regulatory, etc.
- Predictions should **not be randomly** done and a simple change in data/structure can make the predictions horribly wrong
- Several suggested explanation methods (Saliency methods) try to predict **what the model is actually learning** and predicting
- Need to evaluate which saliencey method is the most **useful**
- If **visual doesn't work very well** we need a different type of measures to measure the similarity between saliency maps - mere coincidence

Related Work

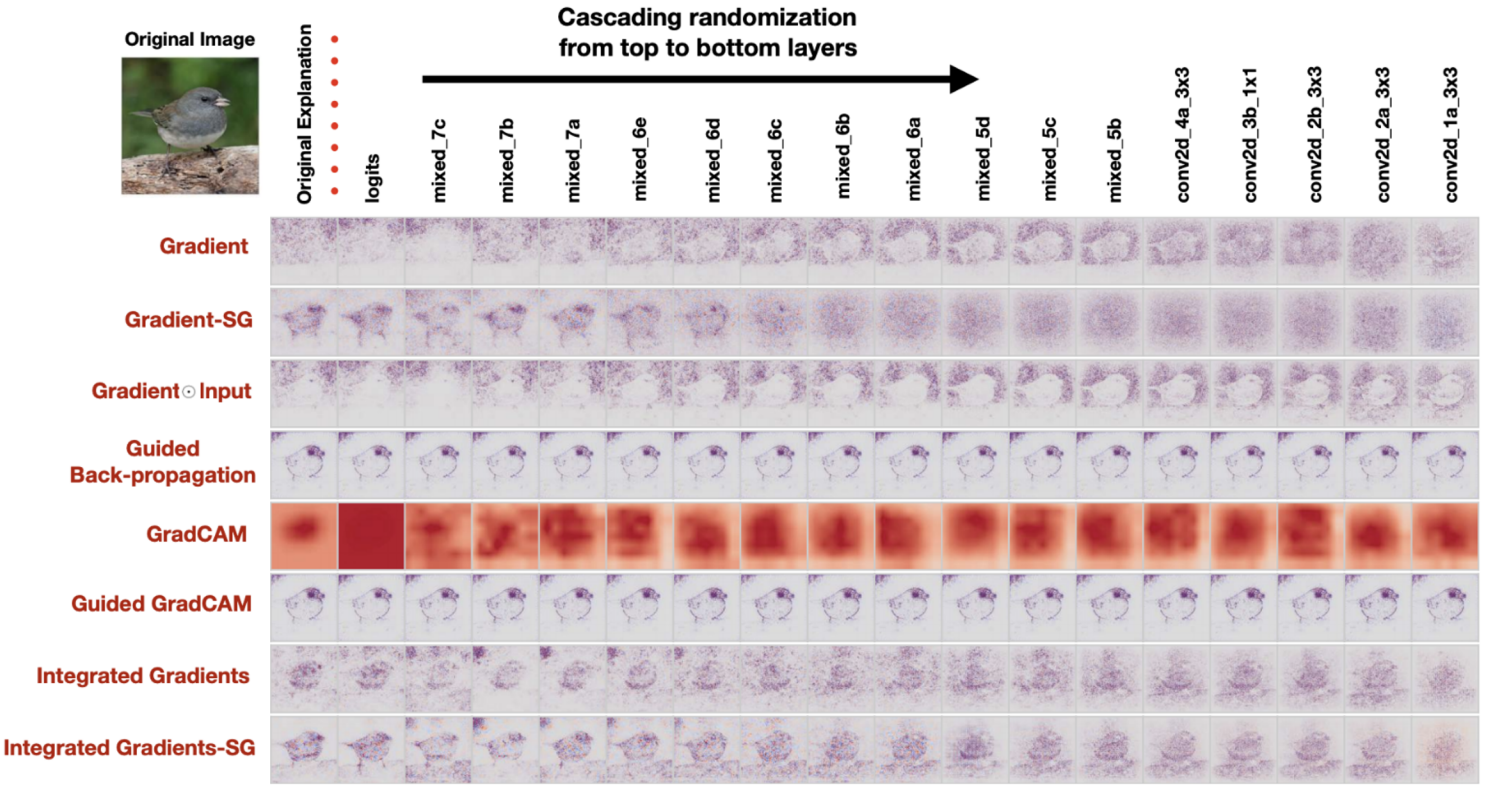
Not a lot I have read but the few I have glossed over:

- Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In ESANN, volume 12, pages 163–172. Citeseer, 2012.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.
- Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806, 2014

Claim / Target Task

- Present 2 different experiments - one on model parameters and one on data labels
  - **Randomization of model parameters**
    - “for a saliency method to be useful for debugging a model, it ought to be sensitive to model parameters.”
  - **Randomization of labels**
    - “..method insensitive to randomizing labels cannot possibly explain mechanisms that depend on the relationship between instances and labels present in the data generating process”
- Compare the saliency maps - visually and using some metrics

An Intuitive Figure Showing WHY Claim



Proposed Solution

- **Randomization of model parameters**
  - 2 experiments - cascaded randomization and independent (layer-wise)
  - Cascaded - top -> top+(top-1) -> … -> top+(top-1)+…1
  - Independent - top -> top-1 -> top-2 … -> 1
  - Assessed the Saliency maps using:
    - Visual
    - HOG similarity
    - SSIM similarity

- **Randomization of labels**
  - "..method insensitive to randomizing labels cannot possibly explain mechanisms that depend on the relationship between instances and labels present in the data generating process"
    - Assessed using Rank Correlation
    - Visual

Implementation

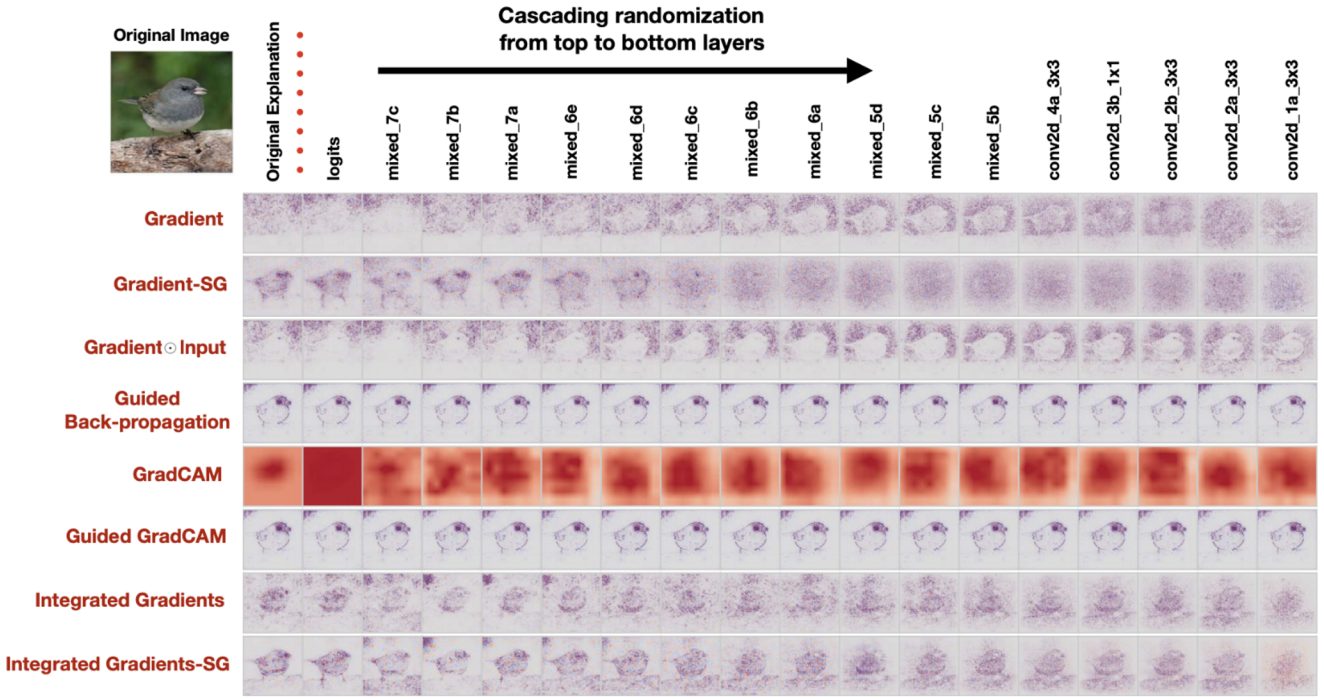
- Trained following models for both experiments on 3 datasets:
  - Imagenet - Inception v3
  - MNIST - MLP
  - Fashion MNIST - CNN

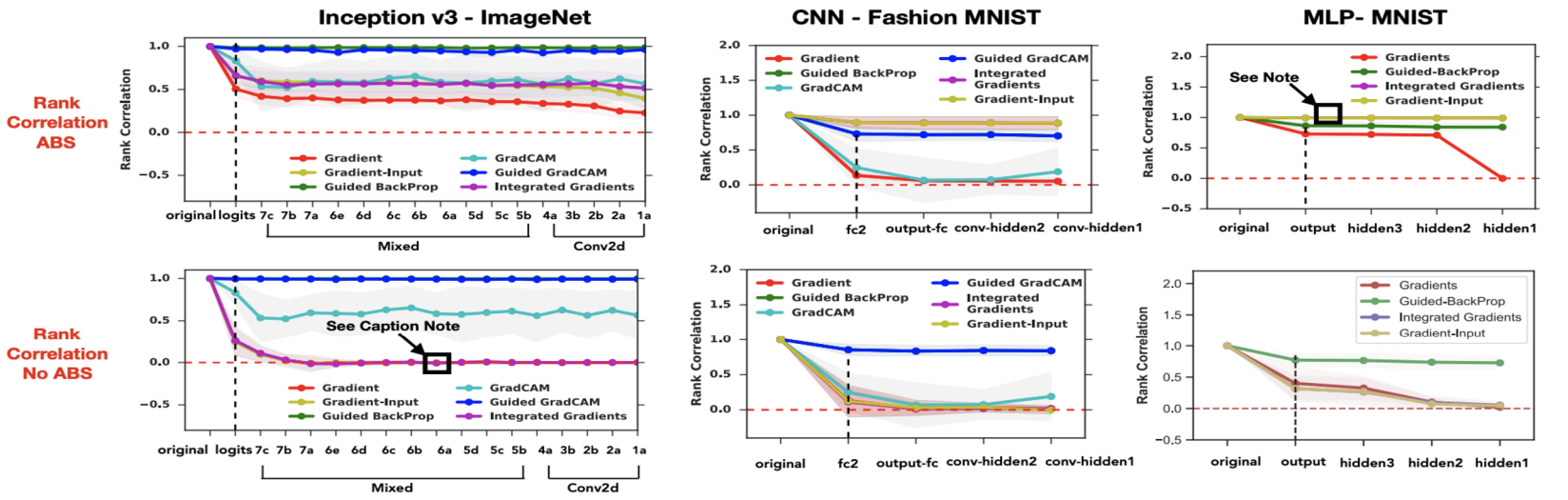
- Methods tested:
  - Gradient
  - Gradient-SG
  - Gradient element-wise Input
  - Guided Backpropagation
  - Guided GradCAM
  - GradCAM
  - Integrated Gradients
  - Integrated Gradients-SG

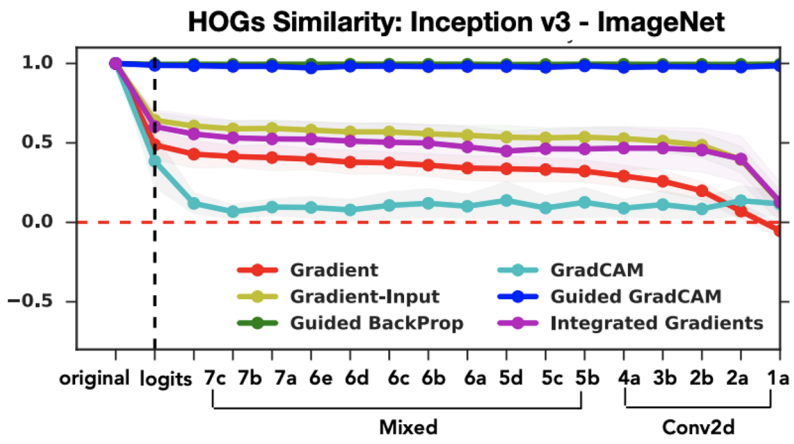
Data Summary

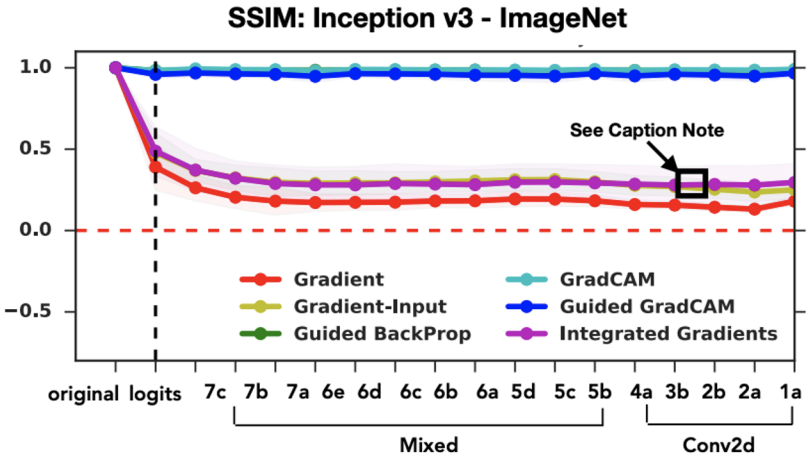
- Imagenet
- MNIST
- Fashion MNIST

Experimental Results - Param

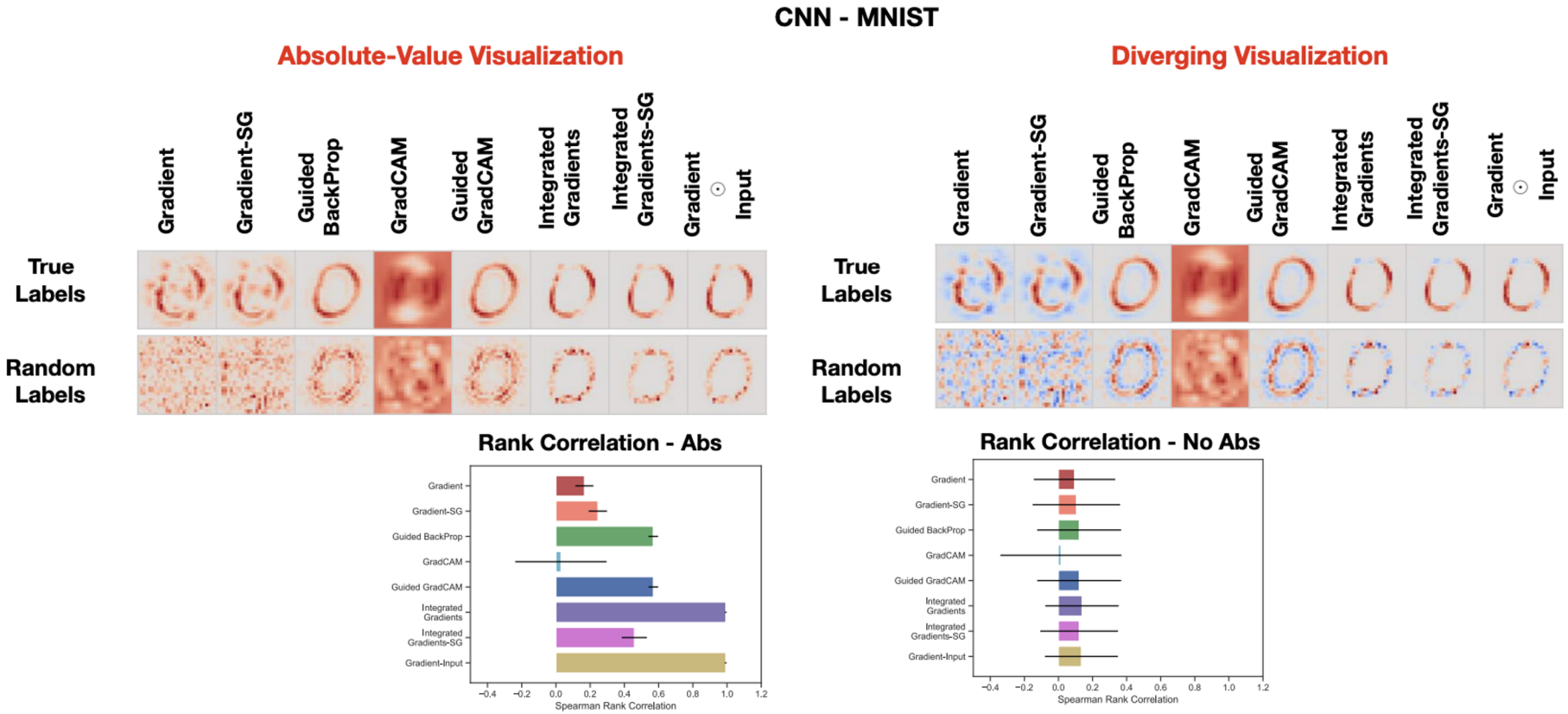



Inception v3 - ImageNet
CNN - Fashion MNIST
MLP- MNIST
Rank Correlation ABS
Rank Correlation No ABS
Rank Correlation
See Caption Note
See Note
Gradient
Gradient-Input
Guided BackProp
GradCAM
Guided GradCAM
Integrated Gradients
Gradients
Guided-BackProp
original logits 7c 7b 7a 6e 6d 6c 6b 6a 5d 5c 5b 4a 3b 2b 2a 1a
Mixed
Conv2d
original fc2 output-fc conv-hidden2 conv-hidden1
original output hidden3 hidden2 hidden1


HOGs Similarity: Inception v3 - ImageNet
Gradient
Gradient-Input
Guided BackProp
GradCAM
Guided GradCAM
Integrated Gradients
original logits 7c 7b 7a 6e 6d 6c 6b 6a 5d 5c 5b 4a 3b 2b 2a 1a
Mixed
Conv2d


SSIM: Inception v3 - ImageNet
See Caption Note
Gradient
Gradient-Input
Guided BackProp
GradCAM
Guided GradCAM
Integrated Gradients
original logits 7c 7b 7a 6e 6d 6c 6b 6a 5d 5c 5b 4a 3b 2b 2a 1a
Mixed
Conv2d

Experimental Results - Labels



Experimental Analysis

Compared the sensitivity among different Saliency methods:

- Guided Backprop and Guided GradCAM (very famous) is broken
- Gradient * Input (element wise) is the most sensitive
- Guided GradCAM no better than an edge detector which is independent of the training data
- There is confirmation bias in where visually equating Saliency maps gives any explanation

Conclusion and Future Work

- Framework for better testing of explanation methods
- Some saliency methods are totally useless

References

- Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In ESANN, volume 12, pages 163–172. Citeseer, 2012.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.
- Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806, 2014