Sanity Checks for 'Saliency' Maps

Julius Adebayo
PhD Student, MIT.

Joint work with





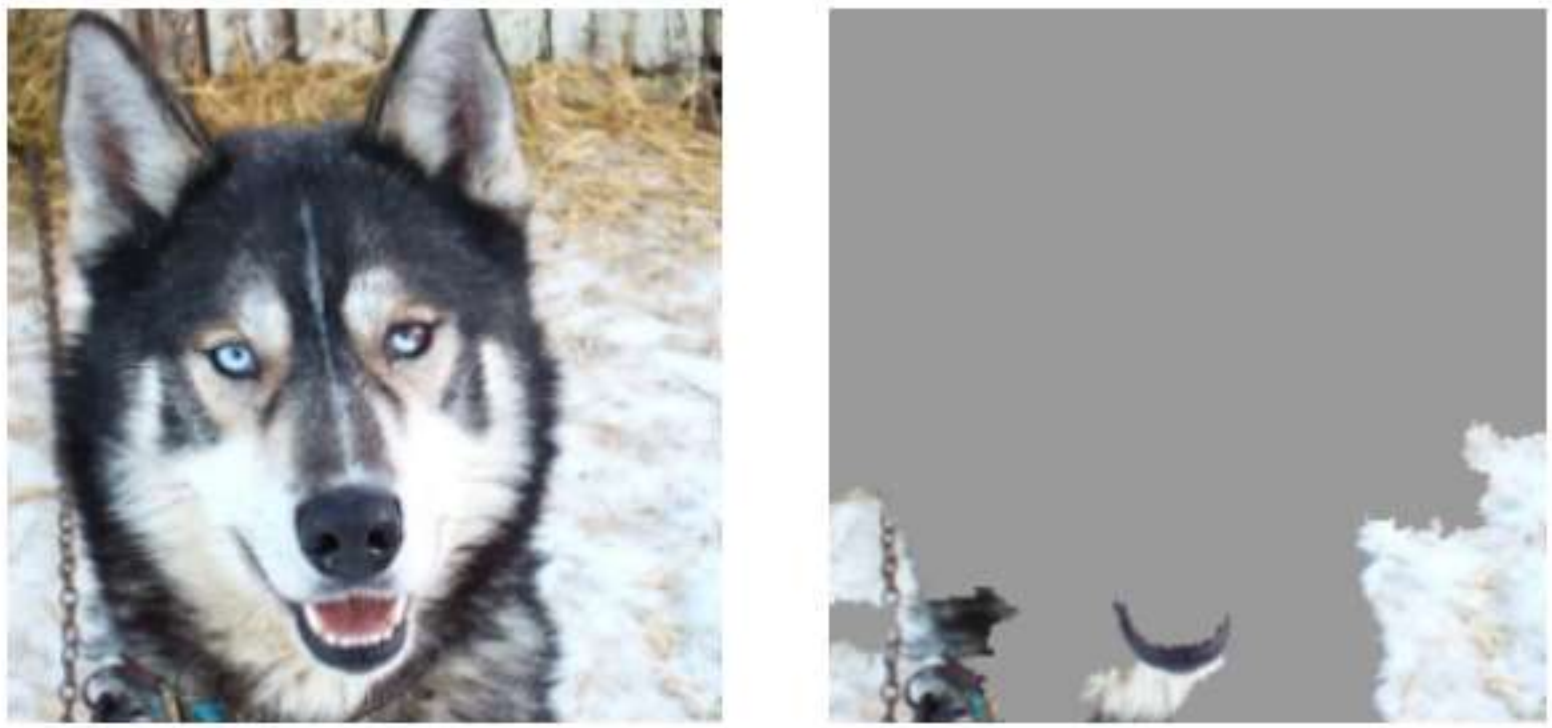
Some Motivation

[Challenges for Transparency, Weller 2017, & Doshi-Velez & Kim, 2017]

- Developer/Researcher: Model Debugging.
- Safety concerns.
- Ethical concerns.
- Trust: Satiate 'societal' need for reasoning to trust an automated system learned from data.

Goals: Model Debugging

- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.

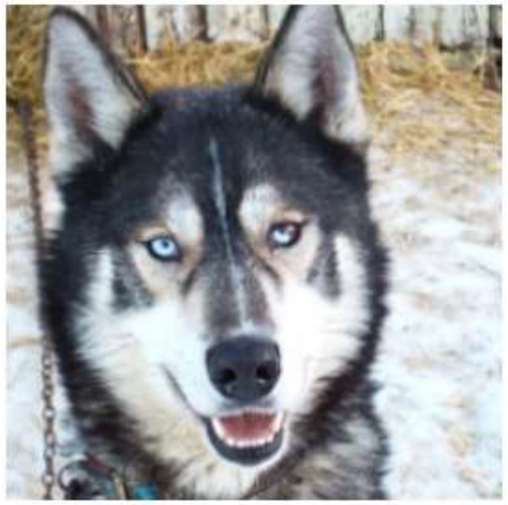


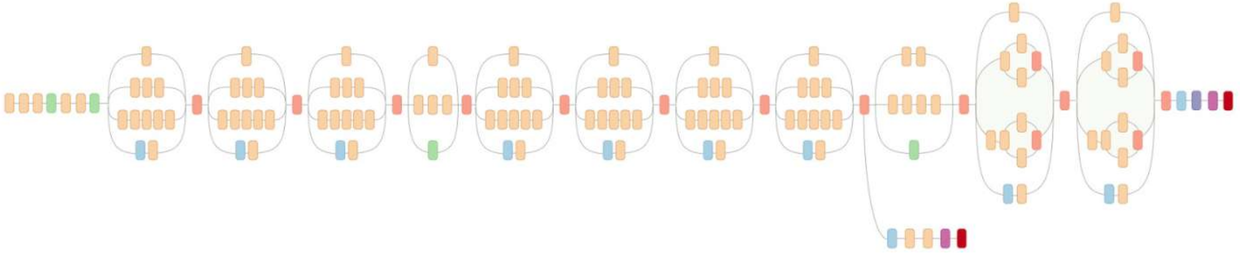
(a) Husky classified as wolf (b) Explanation

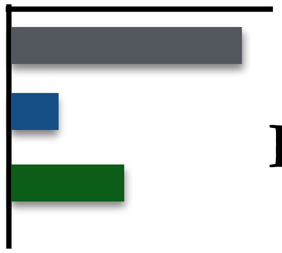
[Ribeiro+ 2016]

Promise of Explanations

- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.

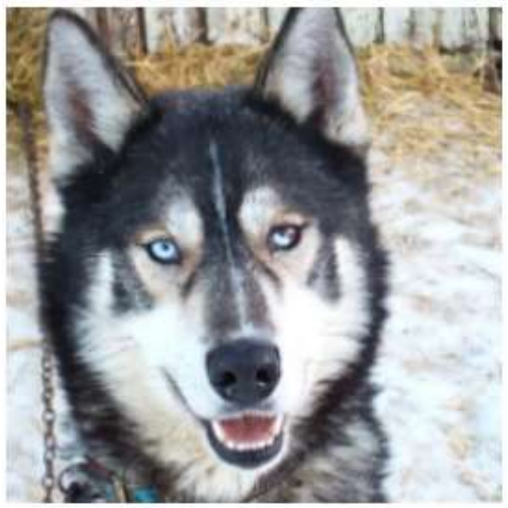


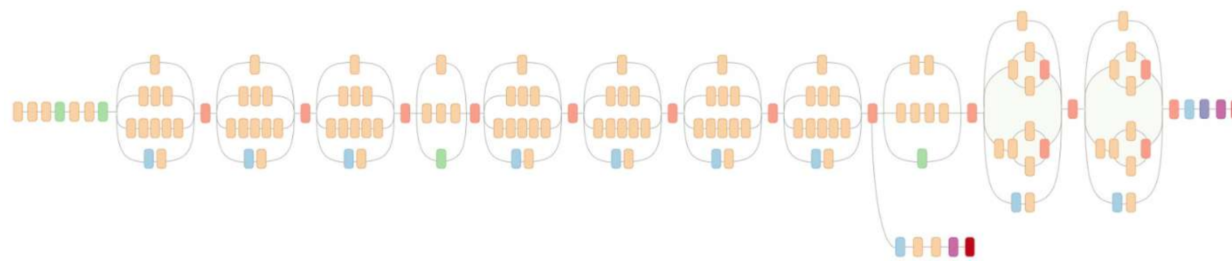


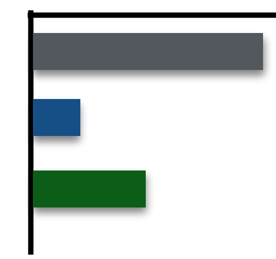


Promise of Explanations

- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.

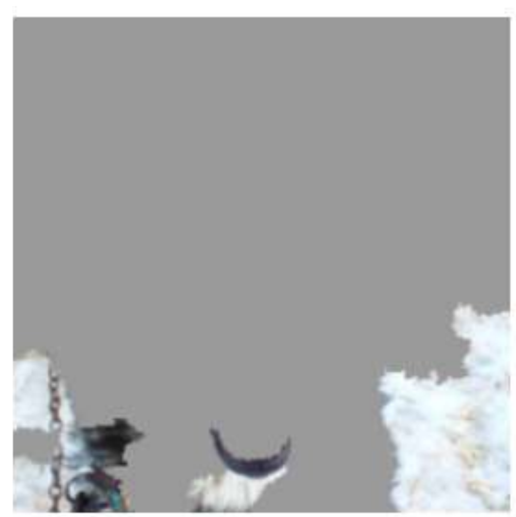






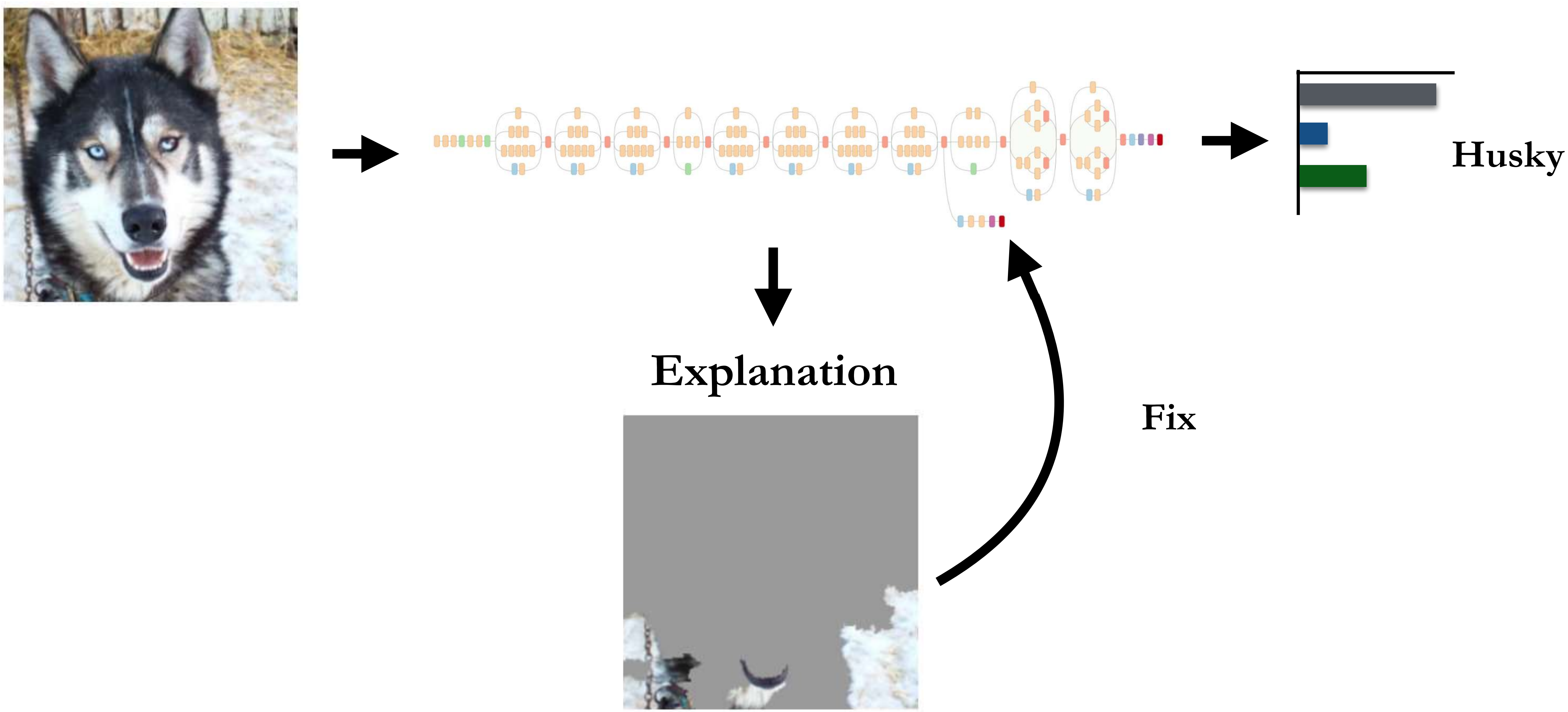
Husky

Explanation



Promise of Explanations

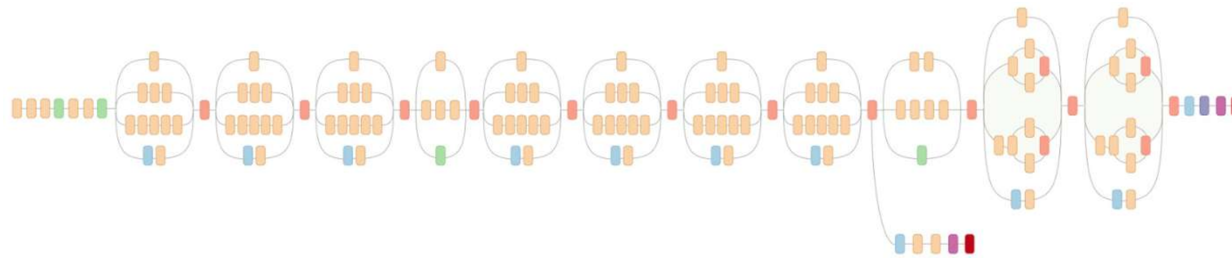
- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.



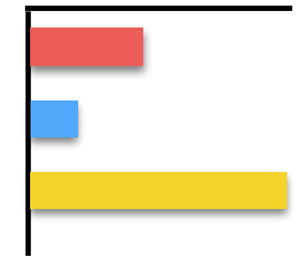
Agenda

- **Overview of attribution methods**
  - **This talk will mostly focus on post-hoc explanation methods for deep neural networks.**
- **The selection conundrum**
- **Sanity checks & results**
- **Theoretical justification by Nie. et. al. 2018.**
- **Passing sanity checks & recent results**
- **Conclusion**

Saliency/Attribution Maps



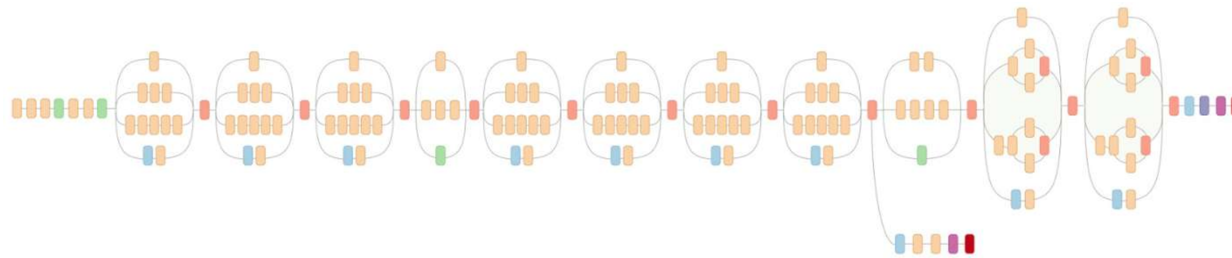
Predictions

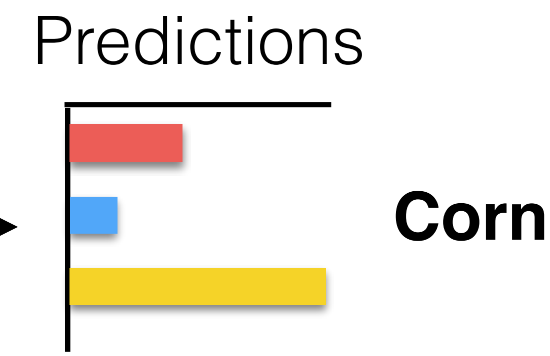


Corn

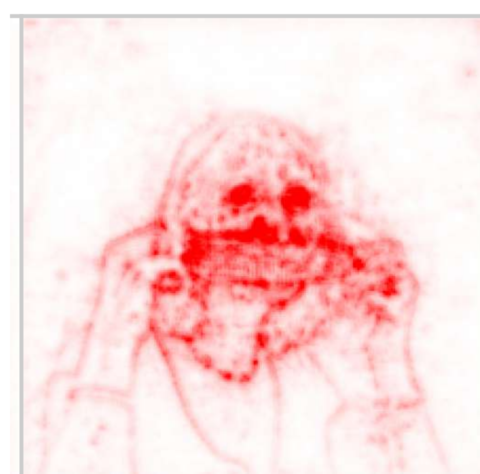
Explanation

Saliency/Attribution Maps



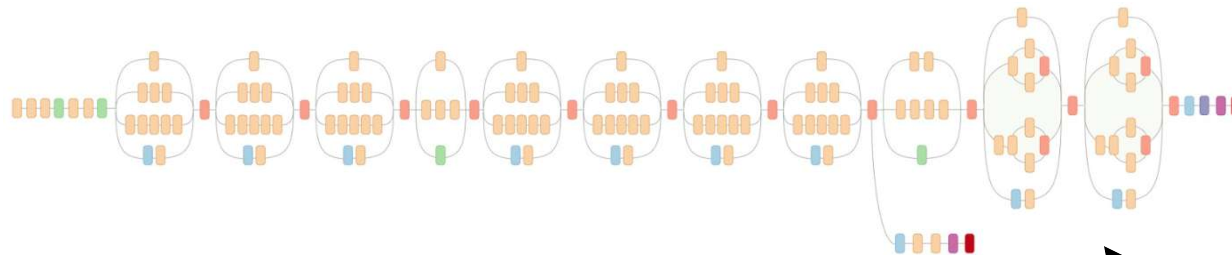


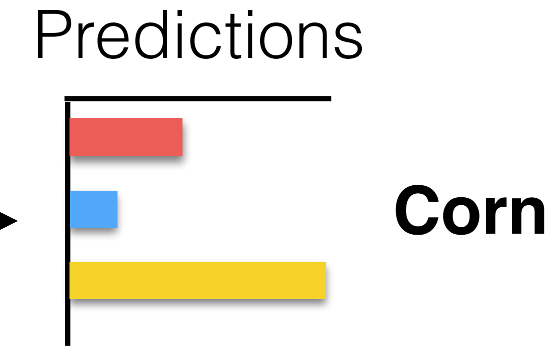
Explanation



Attribution maps provide 'relevance' scores for each dimension of the input.

Saliency/Attribution Maps

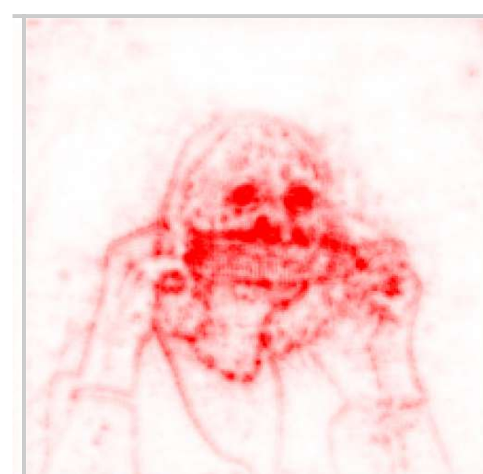




Explanation

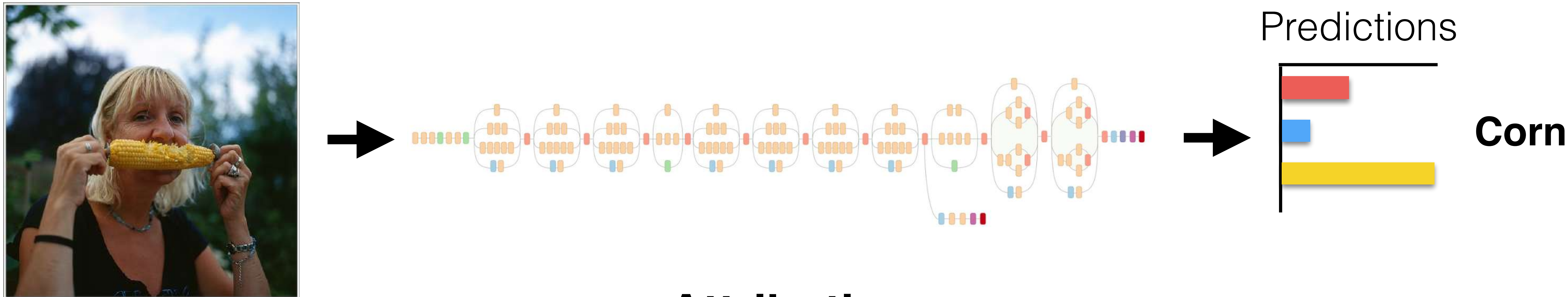
$$S : \mathbb{R}^d \rightarrow \mathbb{R}^C$$

$$E : \mathbb{R}^d \rightarrow \mathbb{R}^d$$



Attribution maps provide 'relevance' scores for each dimension of the input.

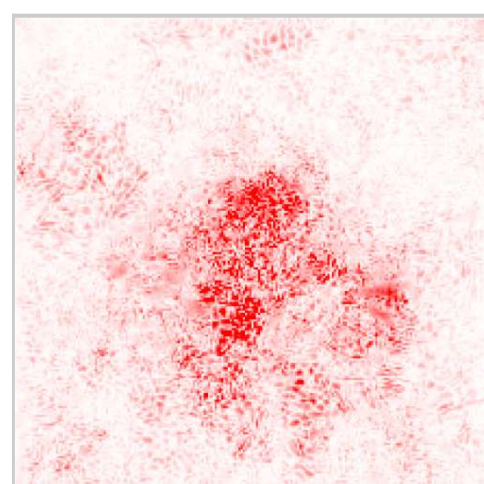
How to compute attribution?



Attribution

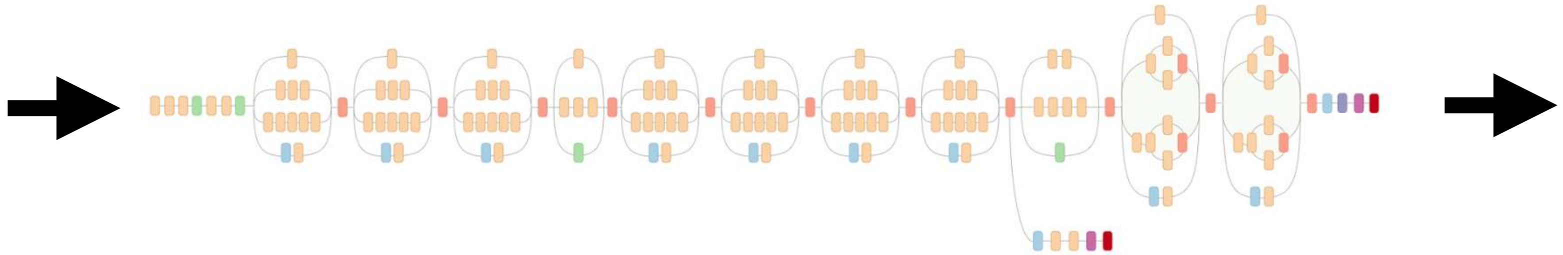
$$E_{grad}(x) = \frac{\partial S_i}{\partial x}$$

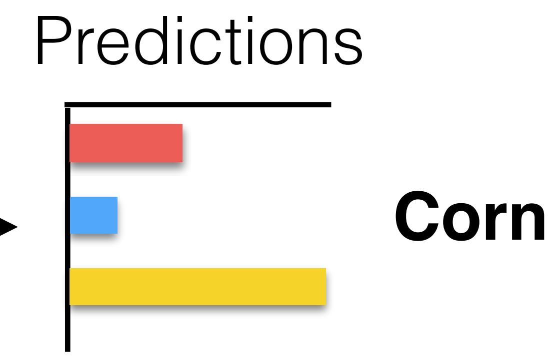
Gradient



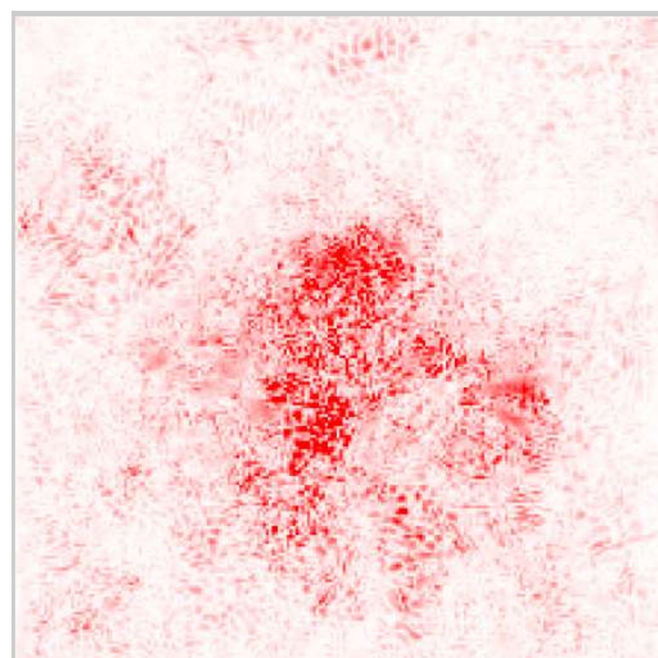
[SVZ'13]

Some Issues with the Gradient



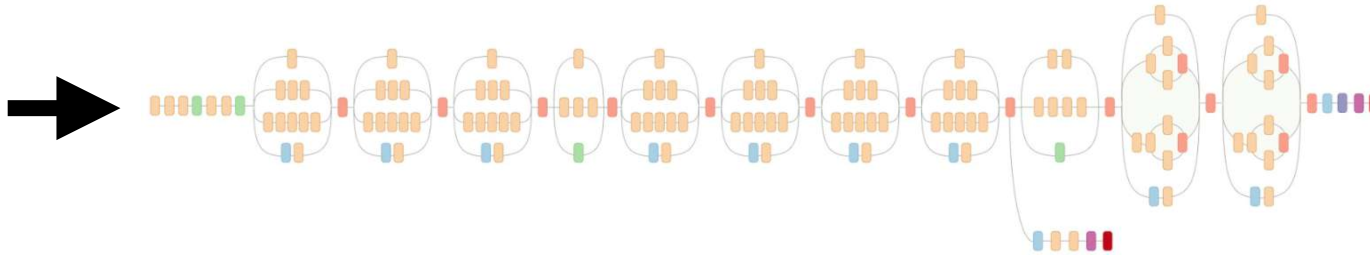


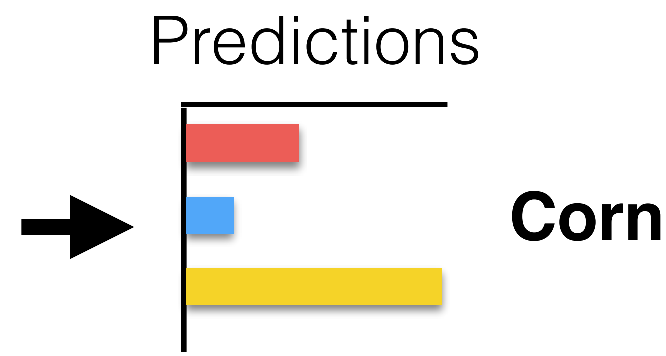
Gradient



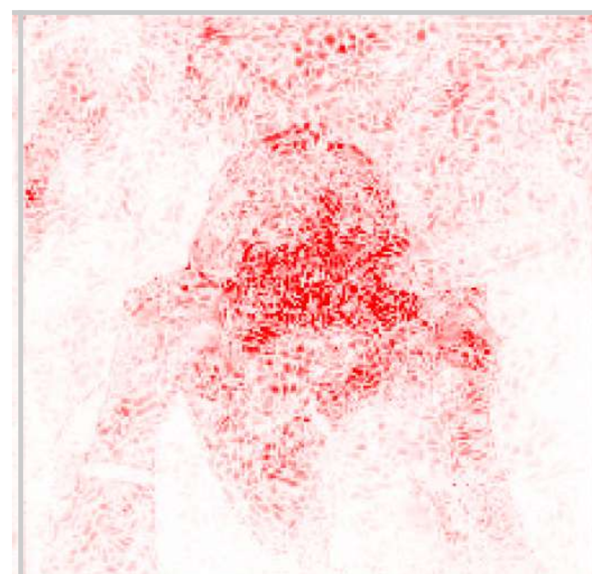
'Visually noisy', and can violate sensitivity w.r.t. a baseline input [Sundararajan et. al., Shrikumar et. al., and Smilkov et. al.]

Integrated Gradients





Integrated Gradients

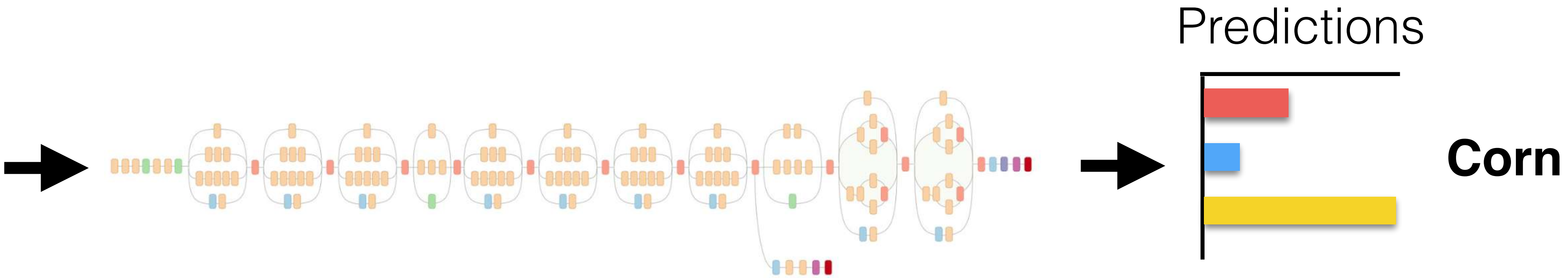


[STY'17]

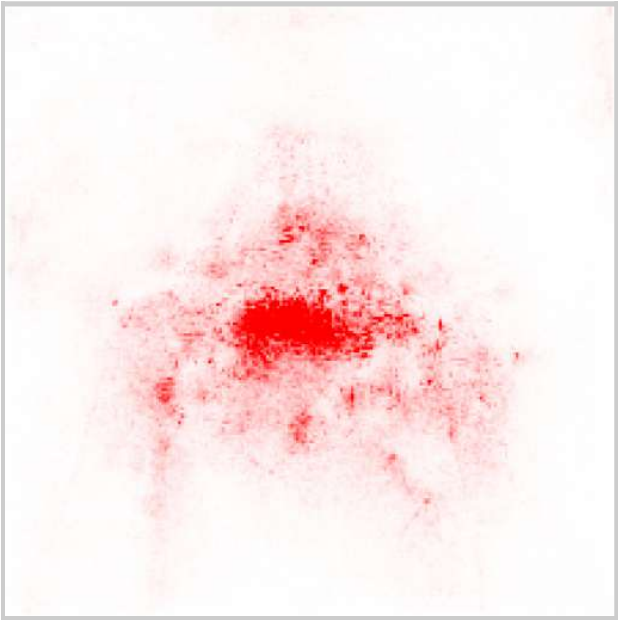
$$E_{\text{IG}}(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial S(\bar{x} + \alpha(x - \bar{x}))}{\partial x} d\alpha$$

Sum of ‘interior’ gradients.

SmoothGrad



SmoothGrad

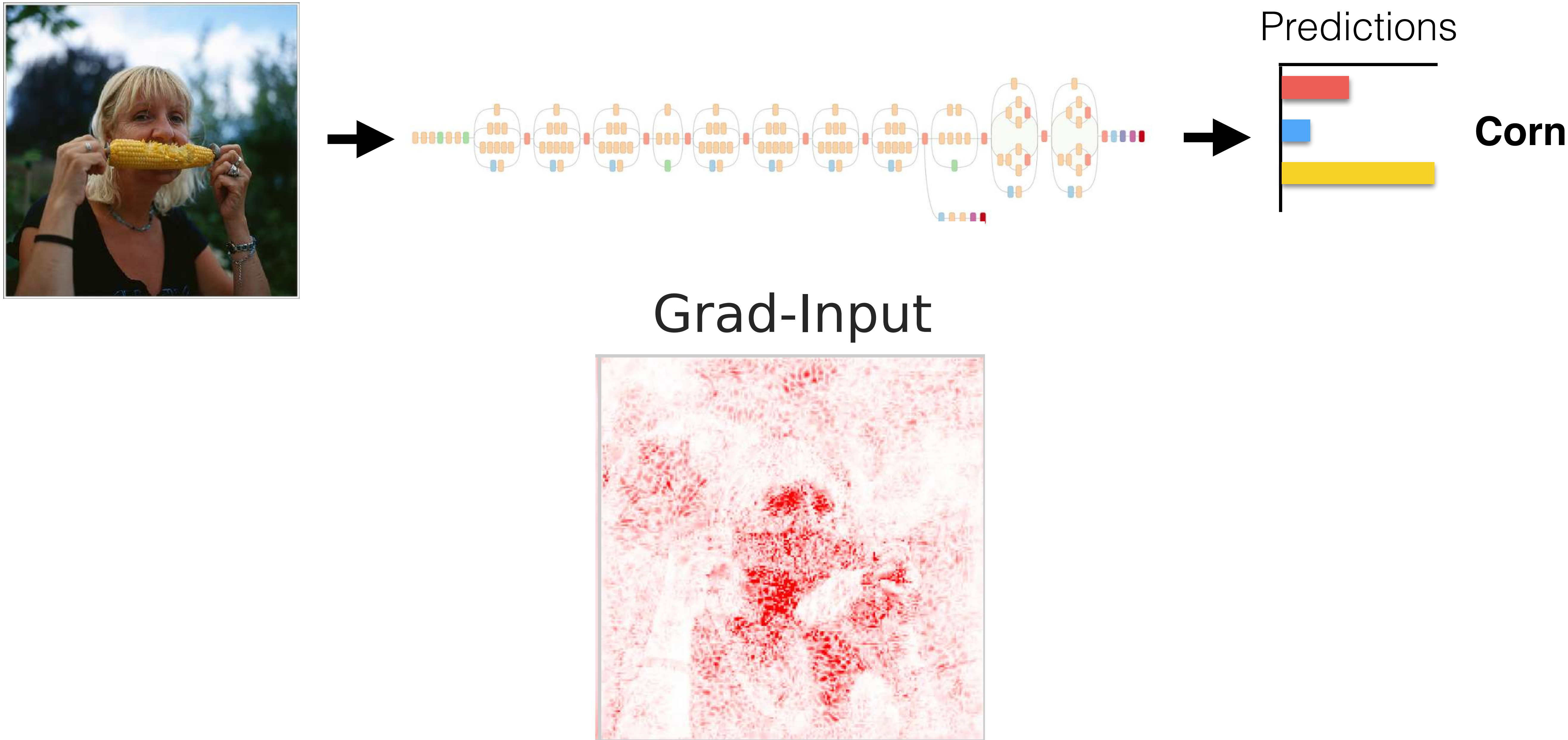


[STKVW’17]

$$E_{\mathrm{sg}}(x) = \frac{1}{N} \sum_{i=1}^{N} E(x + g_i),$$

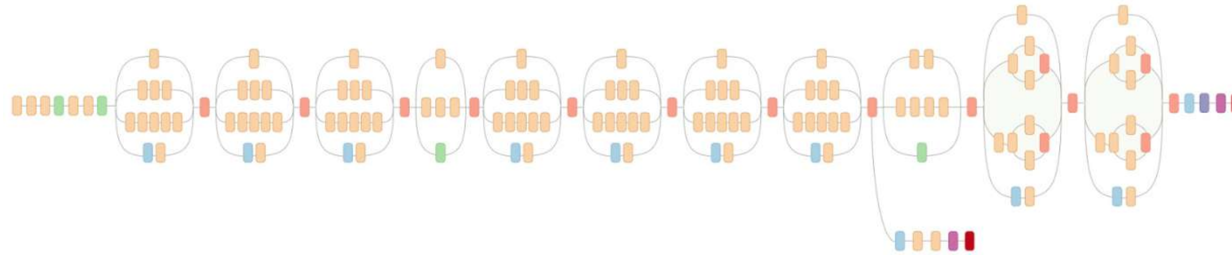
Average attribution of ‘noisy’ inputs.

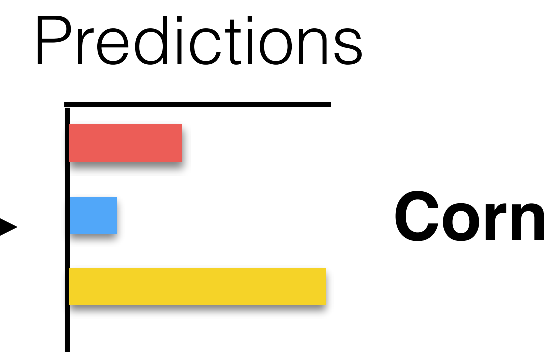
Gradient-Input



Element-wise product of gradient and input.

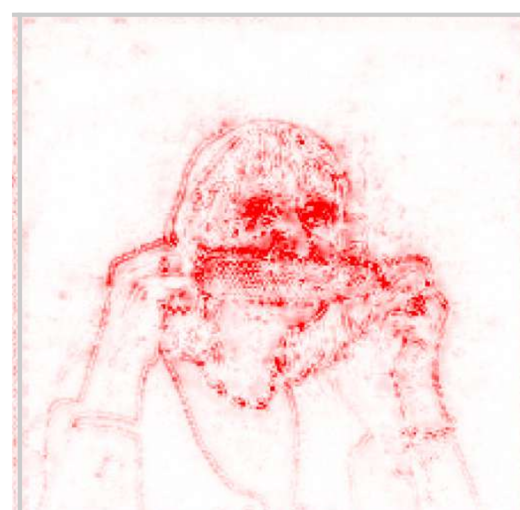
Guided BackProp





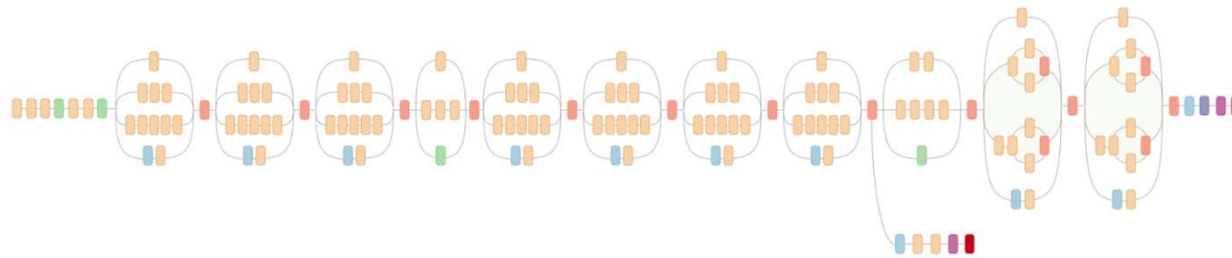
$$\boxed{R^l = 1_{R^{l+1}>0} 1_{f^l>0} R^{l+1}}$$

Guided BackProp

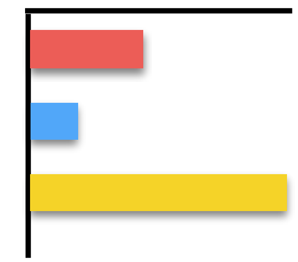


Zero out 'negative' gradients and 'activations' while back-propagating.

Other Learned Kinds



Predictions



Corn

Explanation

[FV'17]

Formulate an explanation as through learned patch removal.

Non-Image Settings: Molecules

Using attribution to decode binding mechanism in neural network models for chemistry

Kevin McCloskey[a,1], Ankur Taly[a,1], Federico Monti[a,b], Michael P. Brenner[a,c], and Lucy J. Colwell[a,d,1]

[a]Google Research, Mountain View, CA 94043; [b]Institute of Computational Science, Università della Svizzera Italiana, CH-6900 Lugano, Switzerland; [c]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [d]Department of Chemistry, Cambridge University, Cambridge CB2 1EW, United Kingdom

Edited by Michael L. Klein, Institute of Computational Molecular Science, Temple University, Philadelphia, PA, and approved April 29, 2019 (received for review December 4, 2018)

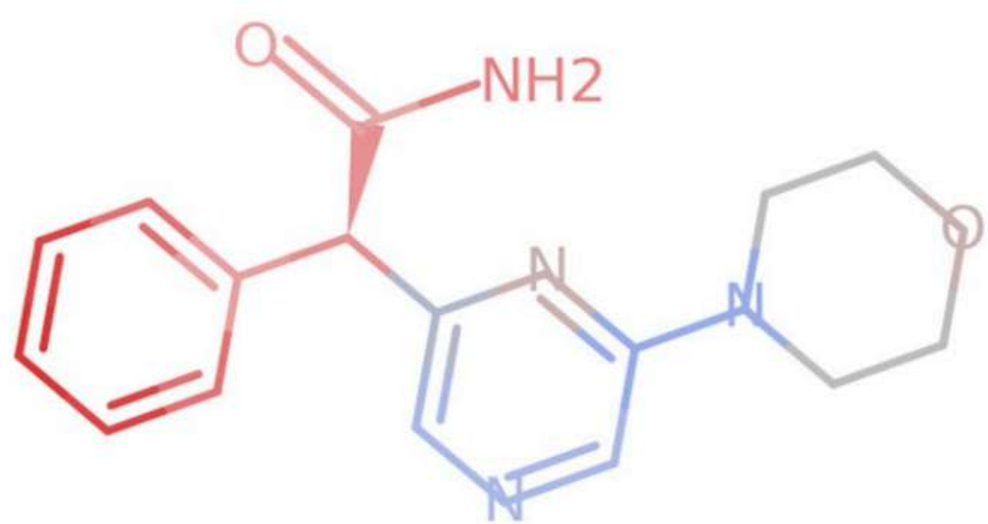
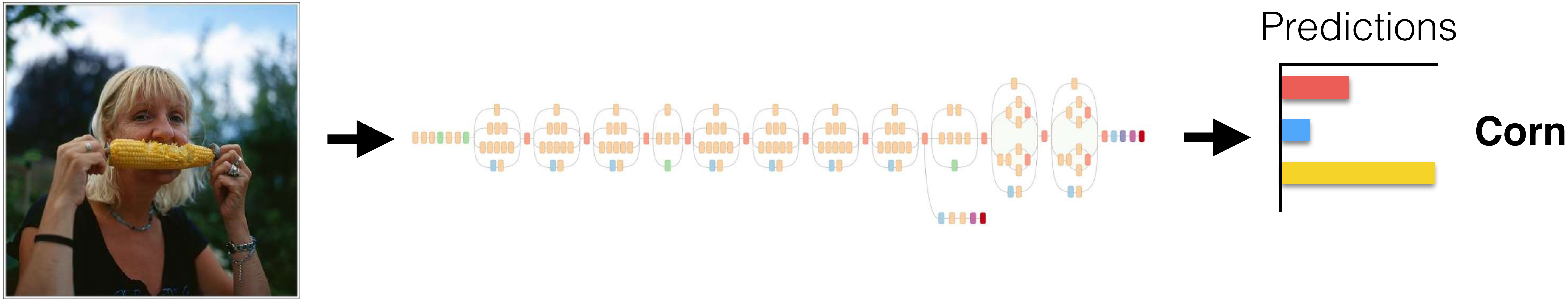
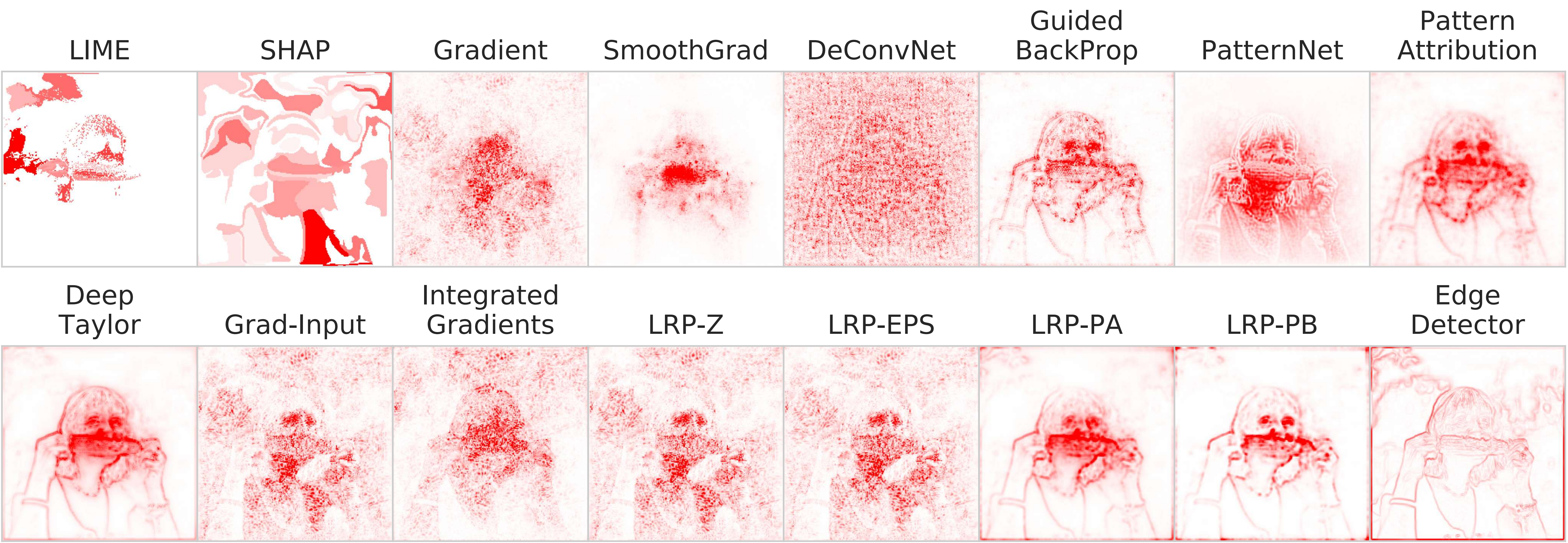


Fig. 1. An example of per-atom model attributions visualized for a molecule. Each atom is colored on a scale from red to blue in proportion to its attribution score, with red being the most positive and blue being the most negative.

The Selection Conundrum





The Selection Conundrum

For a particular task and model, how should a developer/researcher select which method to use?

Desirable Properties

- Sensitivity to the parameters of a model to be explained.
- Depend on the labeling of the data, i.e., reflect the relationship between inputs and outputs.

Sanity Checks

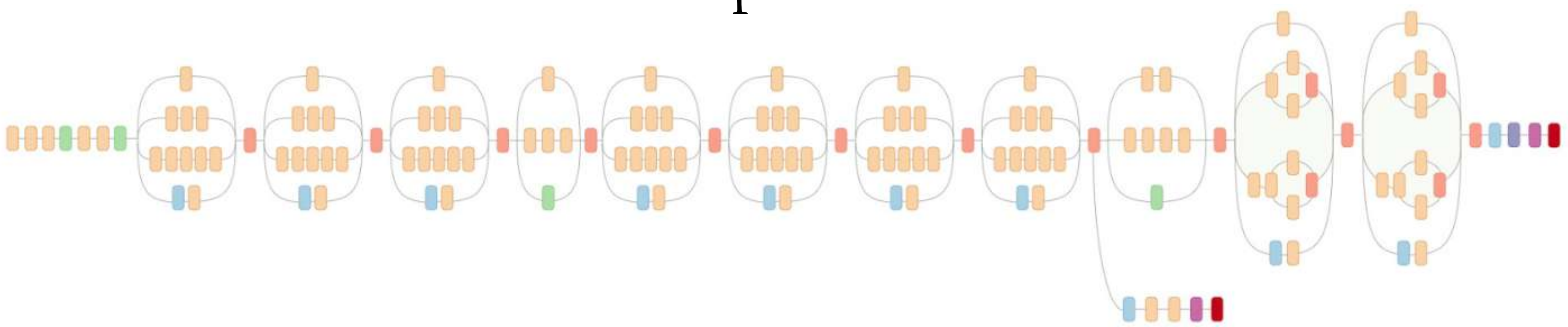
- We will use randomization as a way to test both requirements.

    - **Model parameter randomization test:** randomize (re-initialize) the parameters of a model and now compare attribution maps for a trained model to those derived from a randomized model.

    - **Data randomization test**: compare attribution maps for a model trained with correct labels to those derived from a model trained with random labels.

Model Parameter Randomization

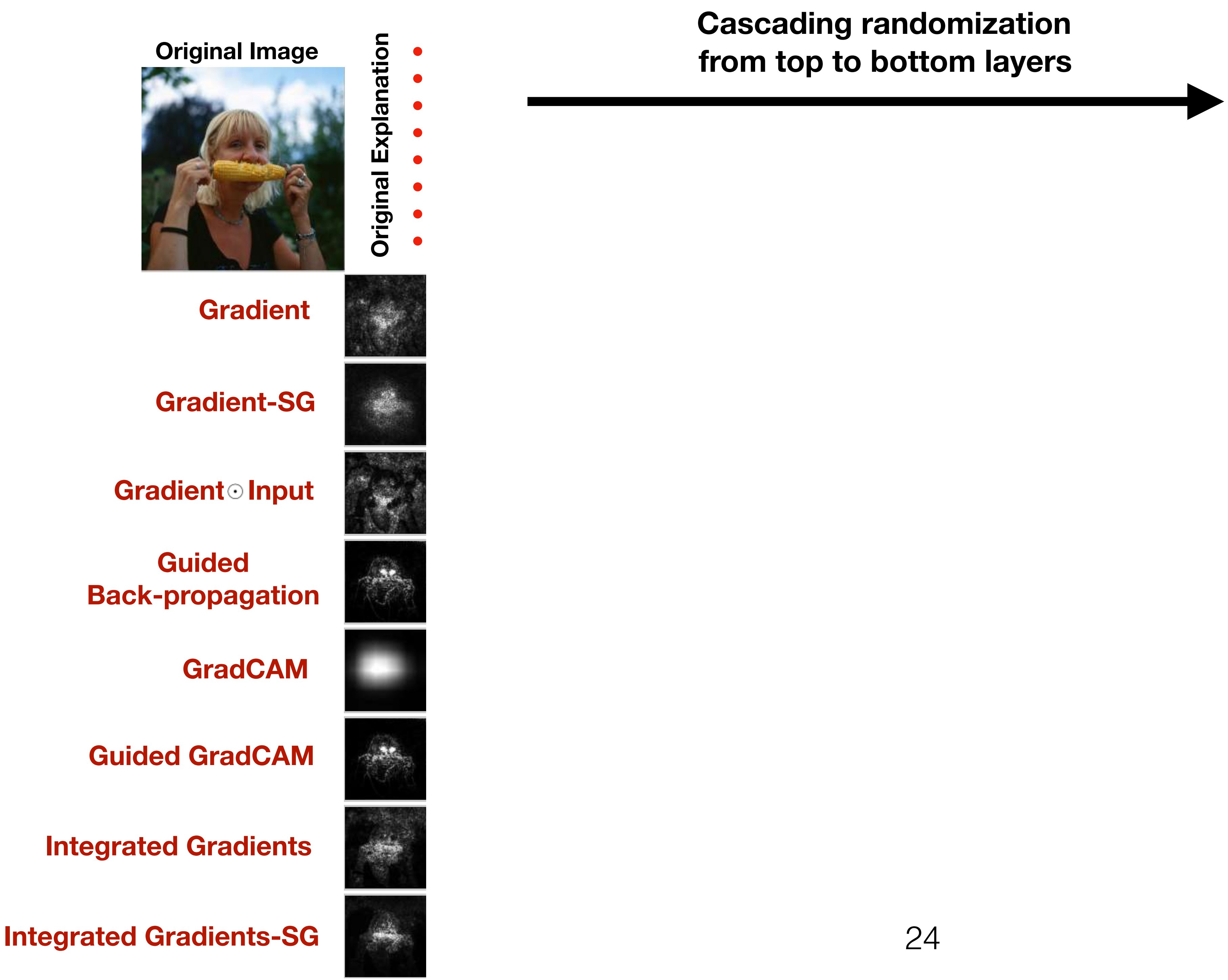
Inception V3



- Cascading randomization from top to bottom layers.
- Independent layer randomization.

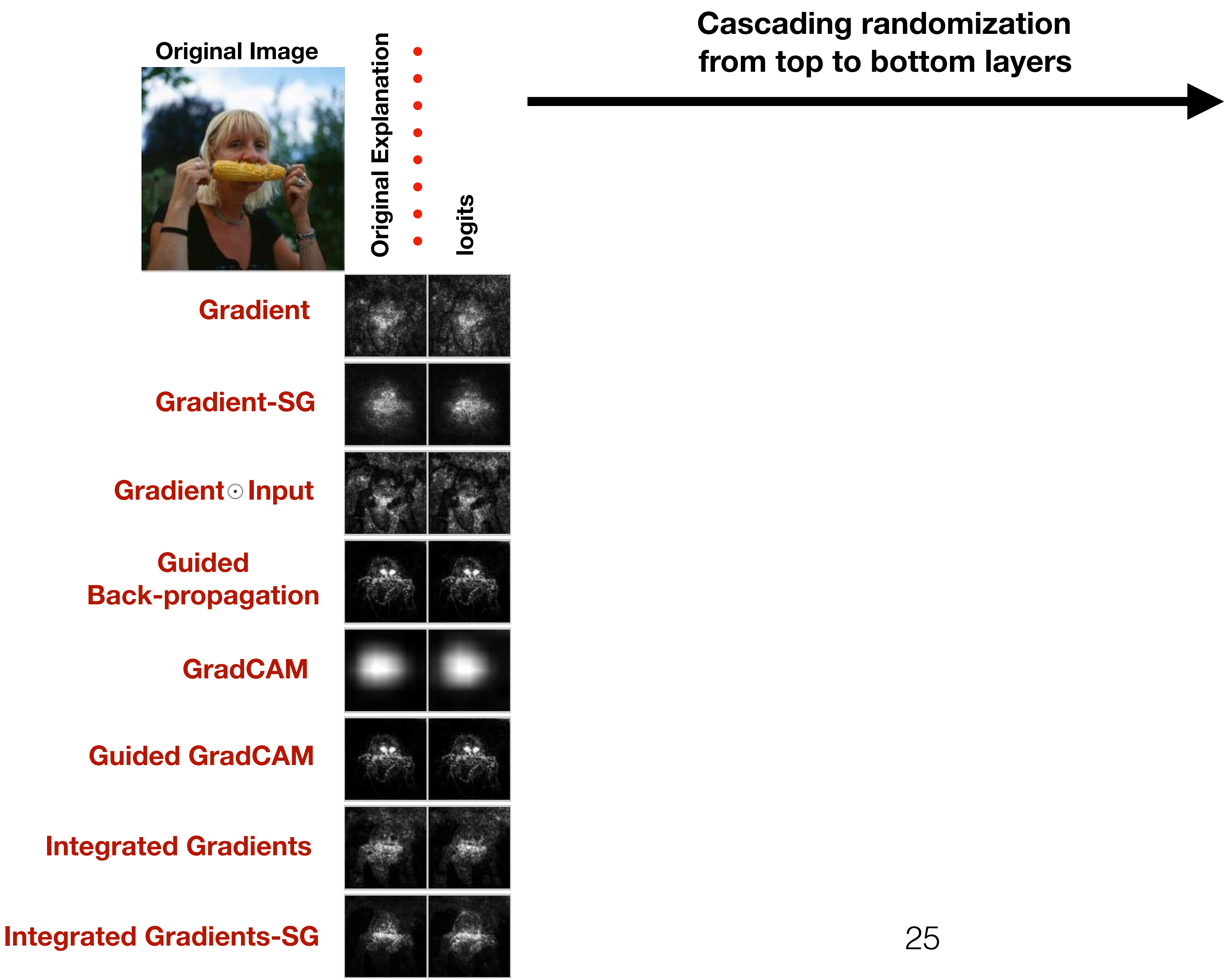
Model Parameter Randomization

Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.



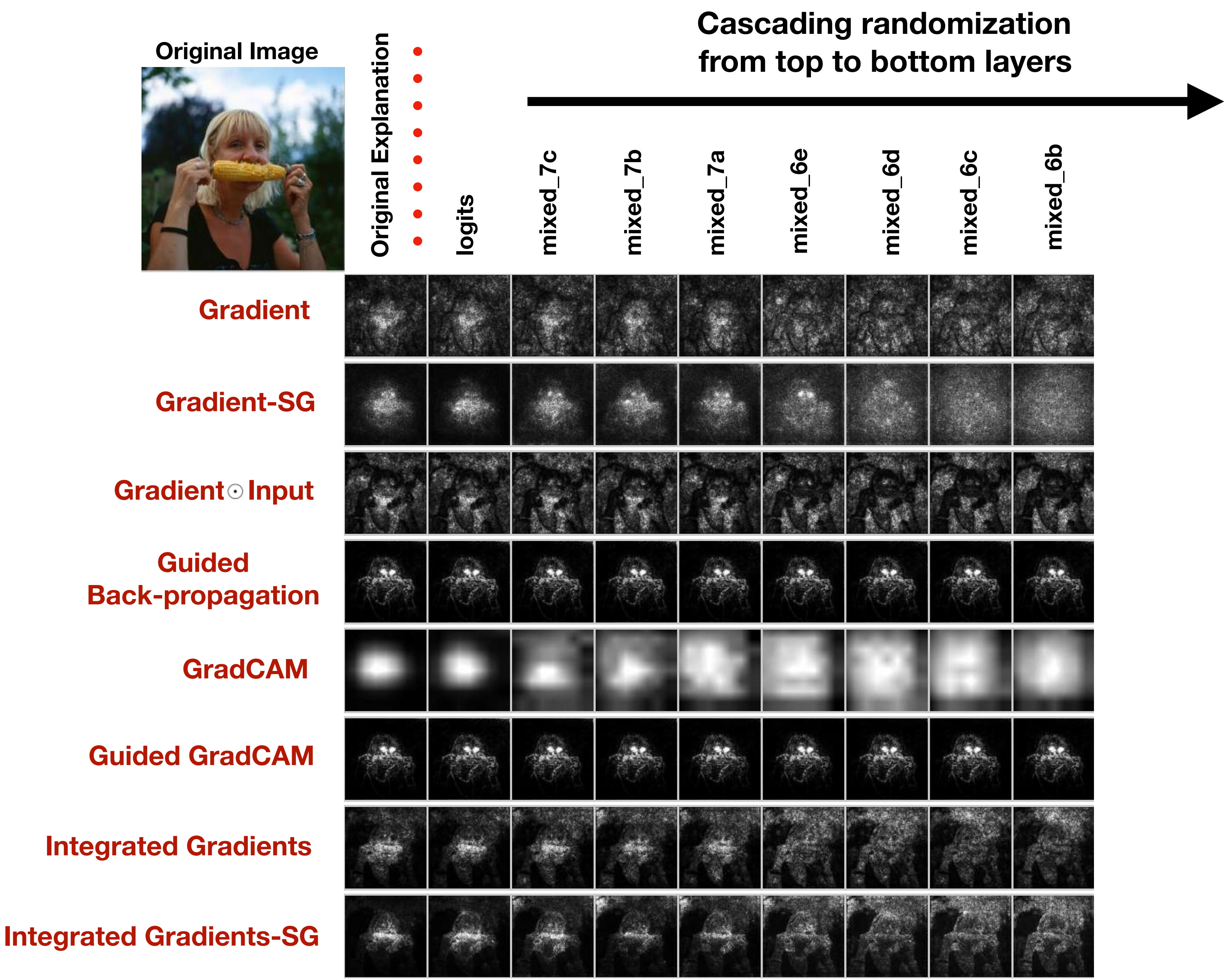
Model Parameter Randomization

Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.



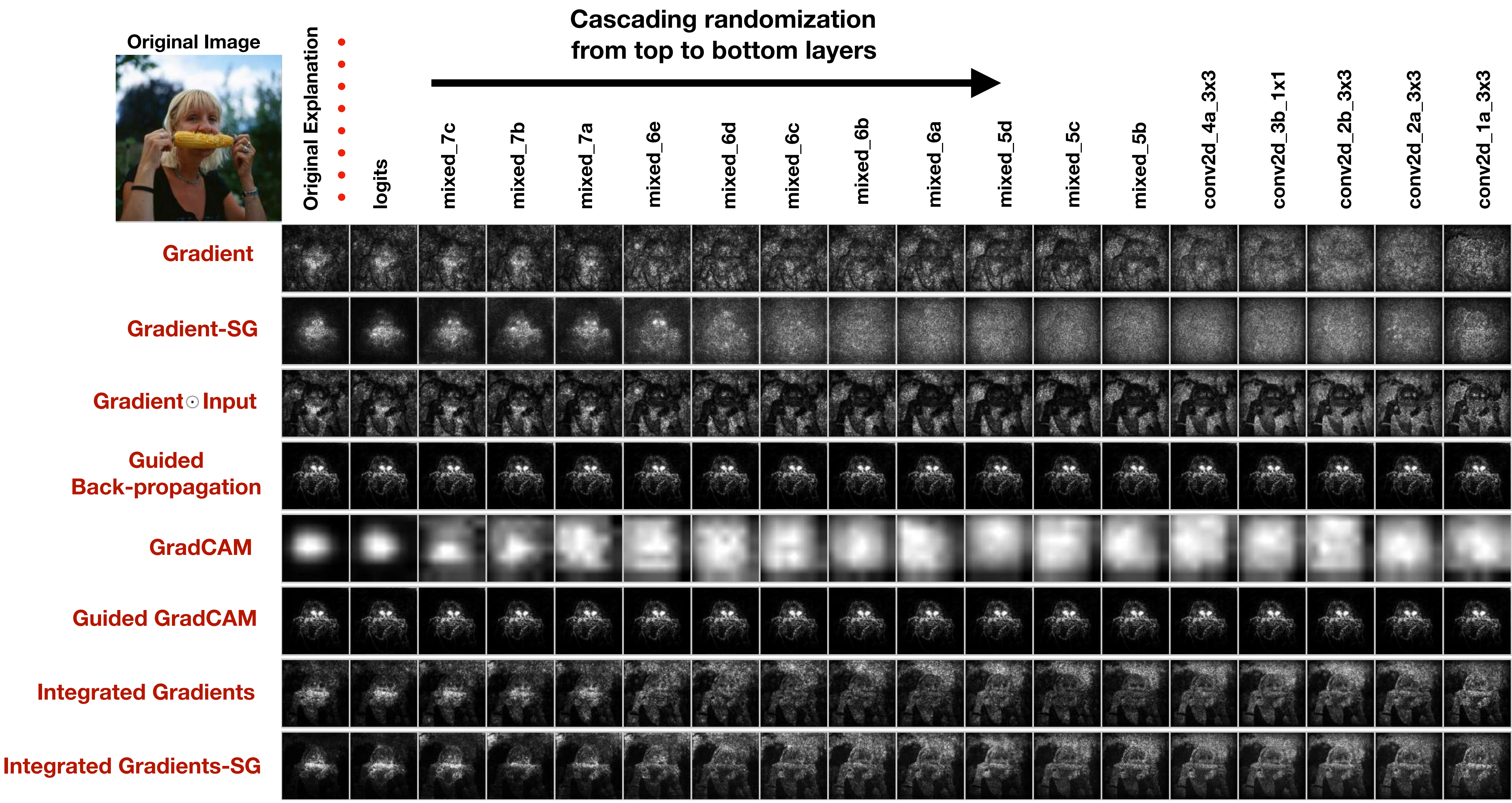
Model Parameter Randomization

Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.



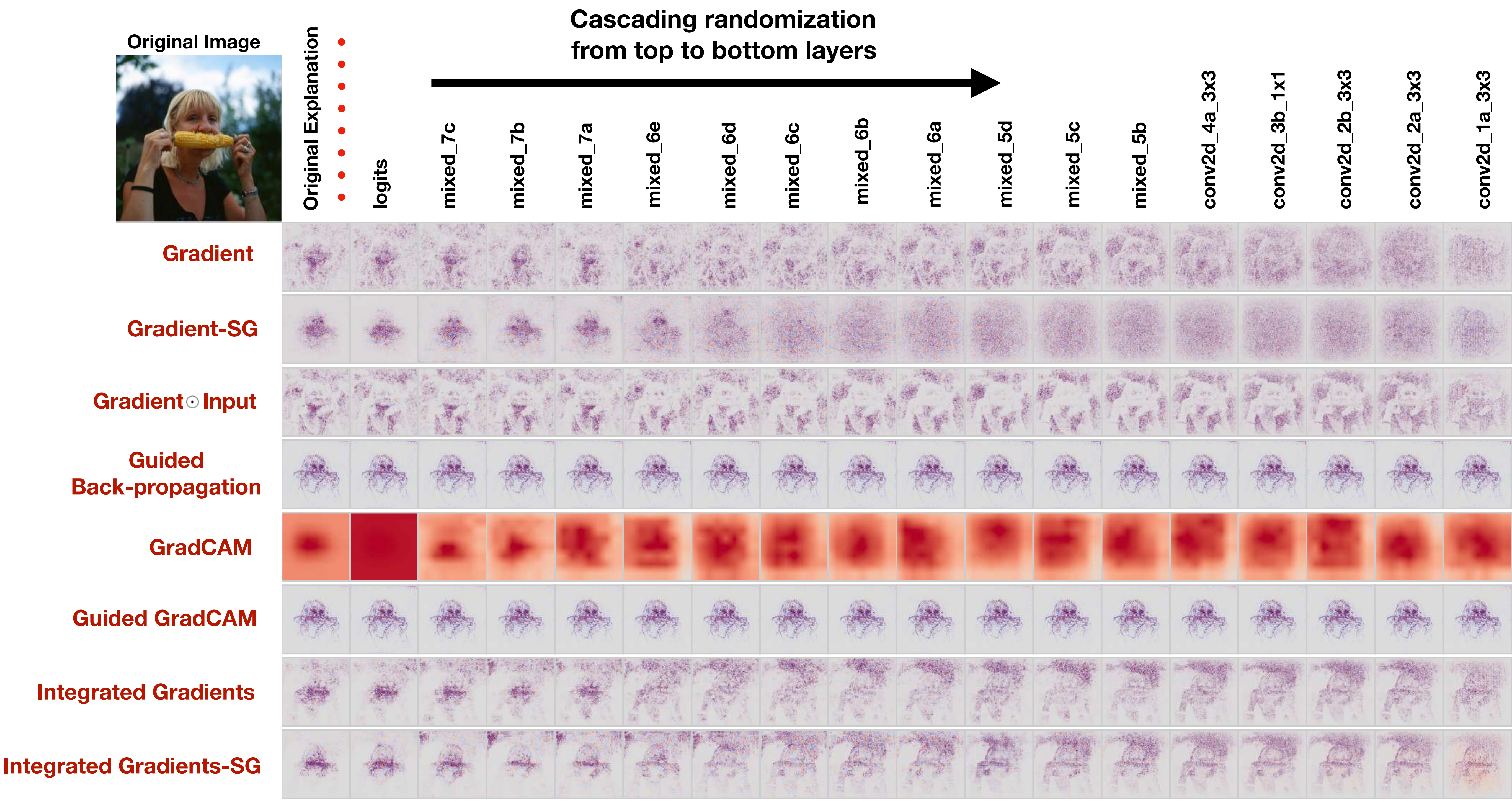
Model Parameter Randomization

Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.



Model Parameter Randomization

Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.

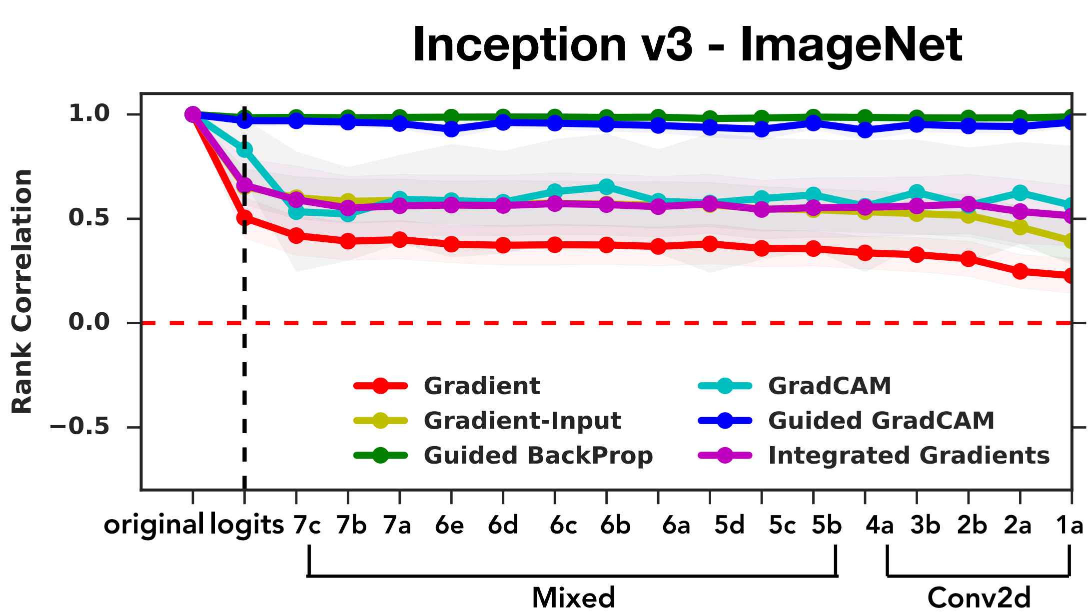


Metrics

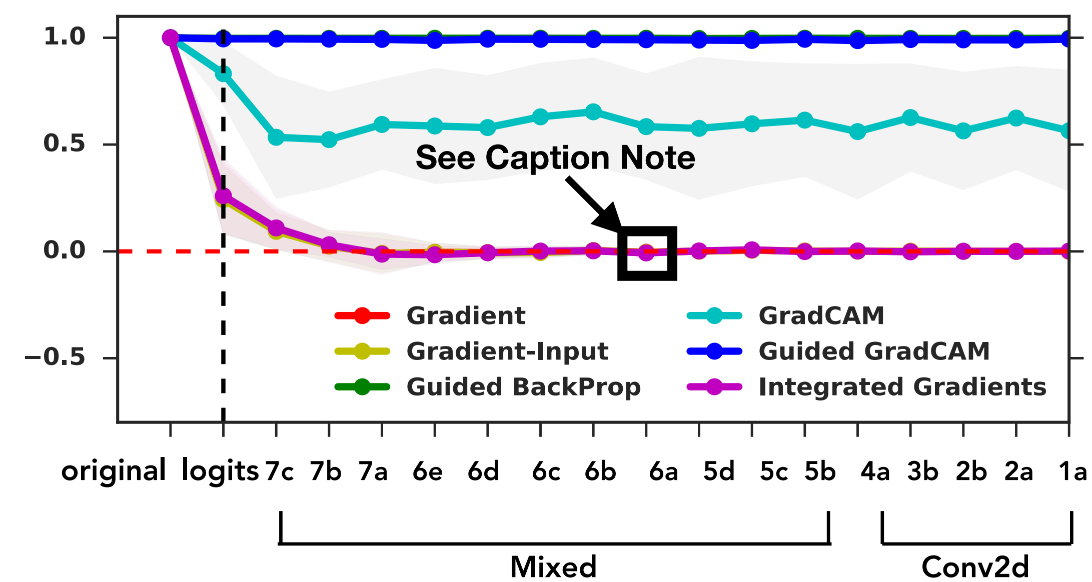
- Rank correlation of attribution from model with trained weights to those derived from partially randomized models.

- Attribution sign changes. Roughly similar regions are, however, still attributed.

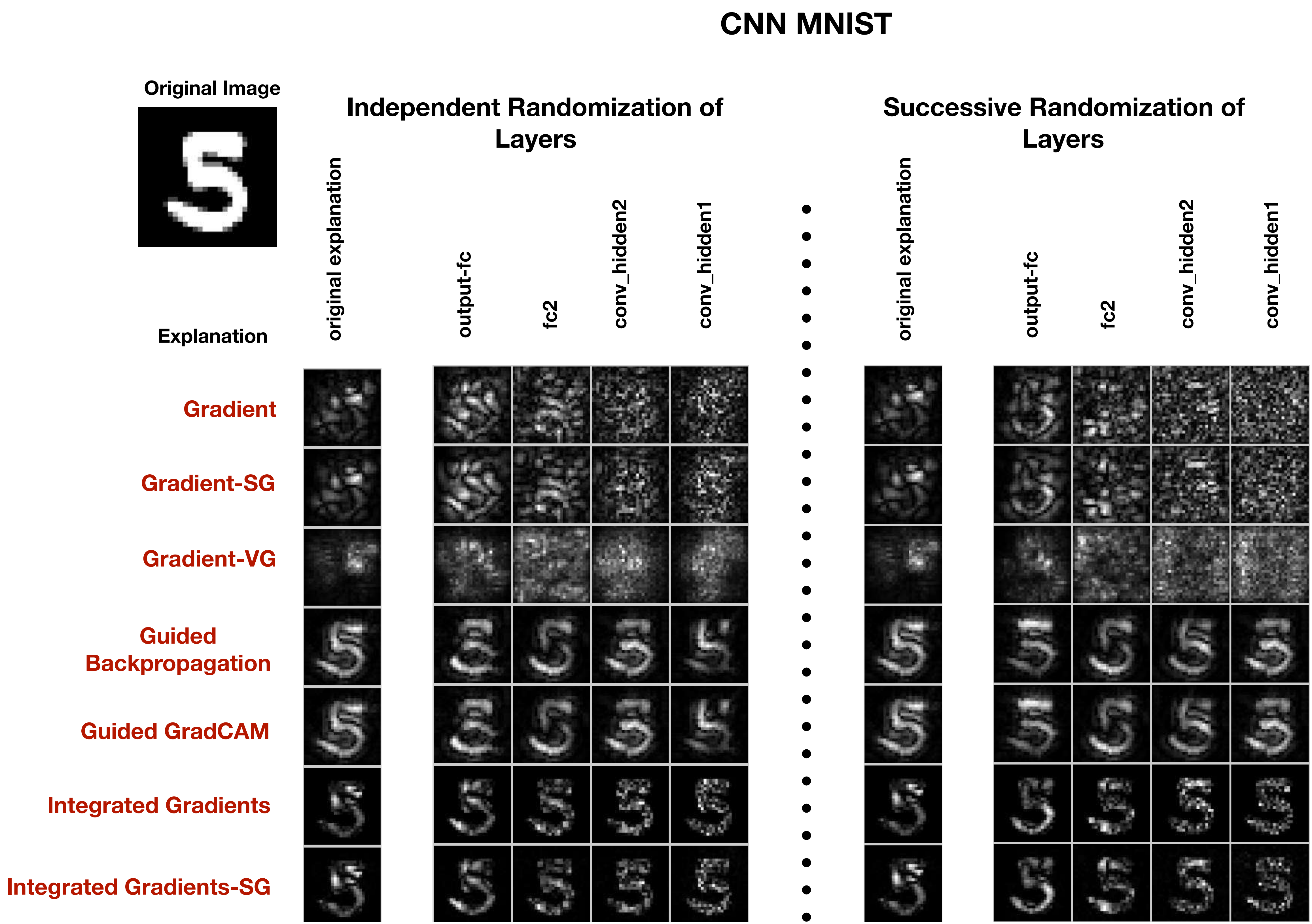
Rank Correlation ABS



Rank Correlation No ABS

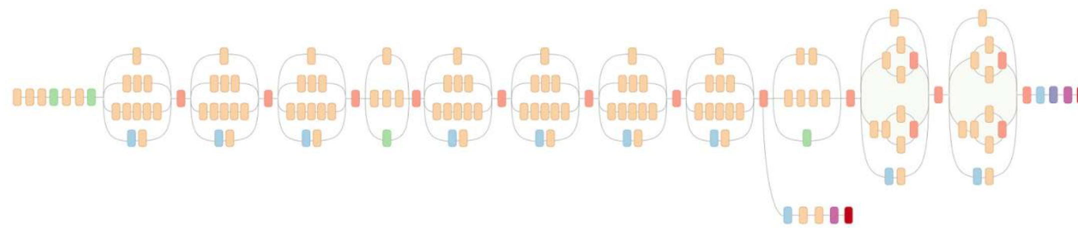


Model Parameter Randomization



Medical Setting

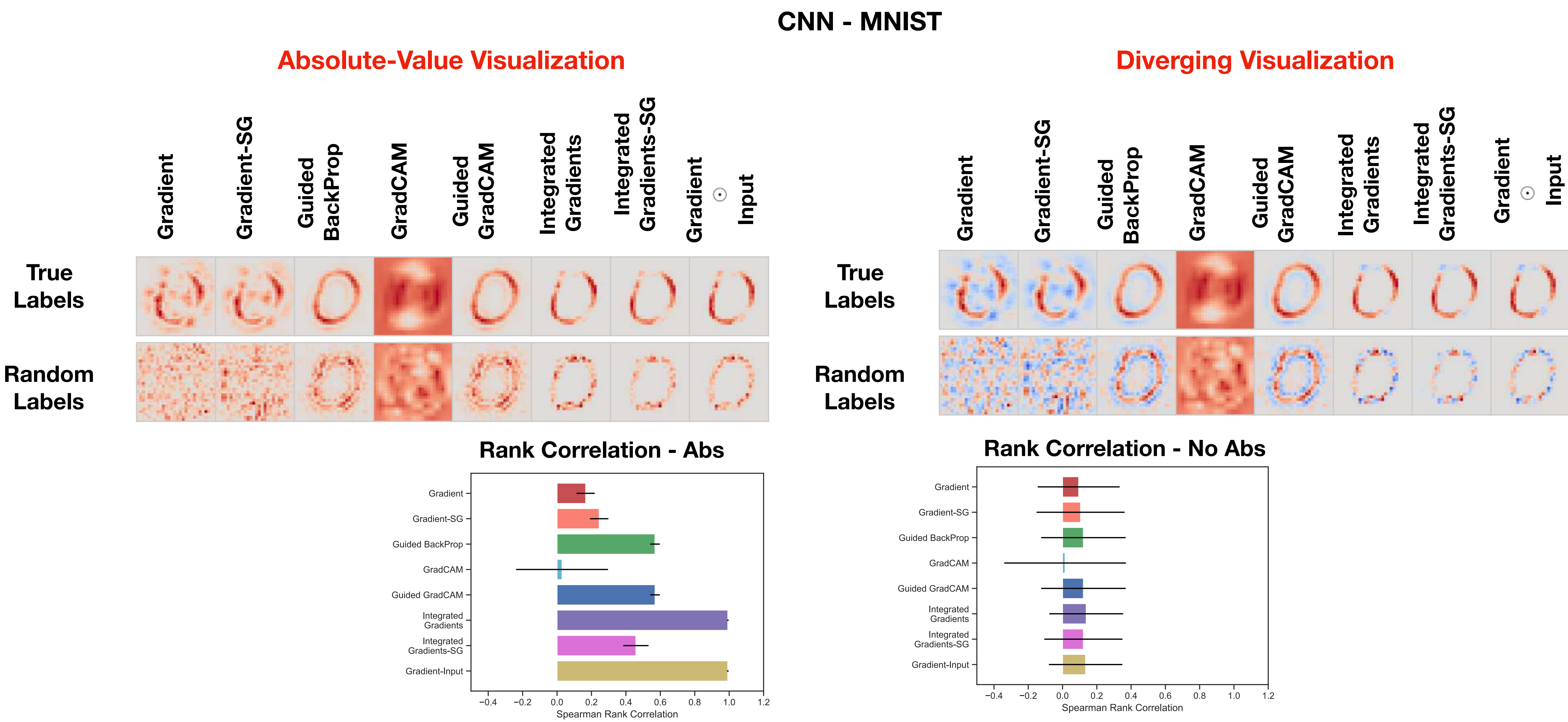
Skeletal Radiograph



Age

Guided Backpropagation

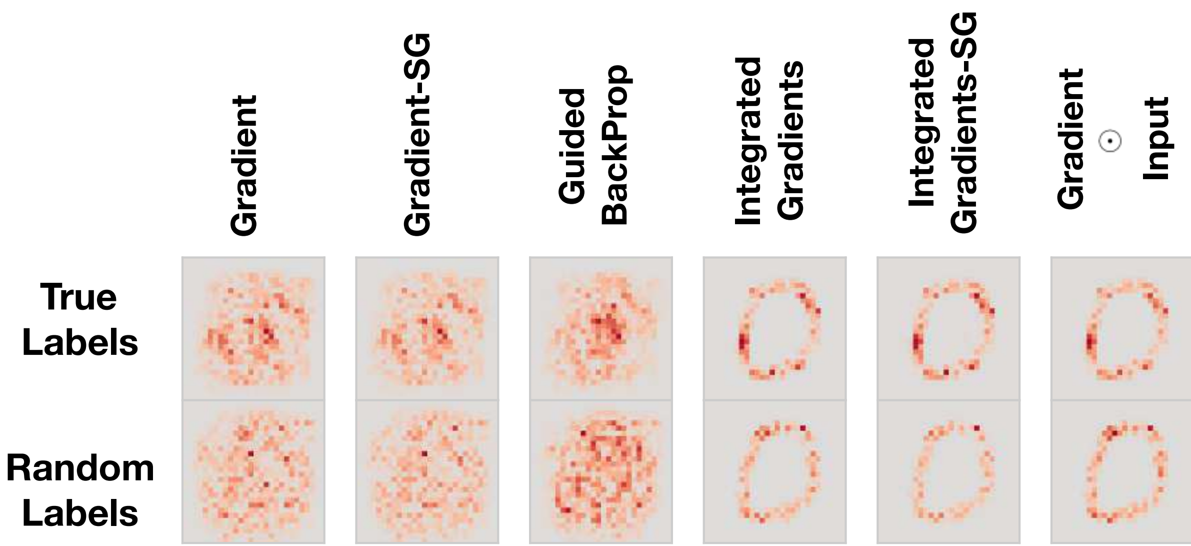
Data Randomization



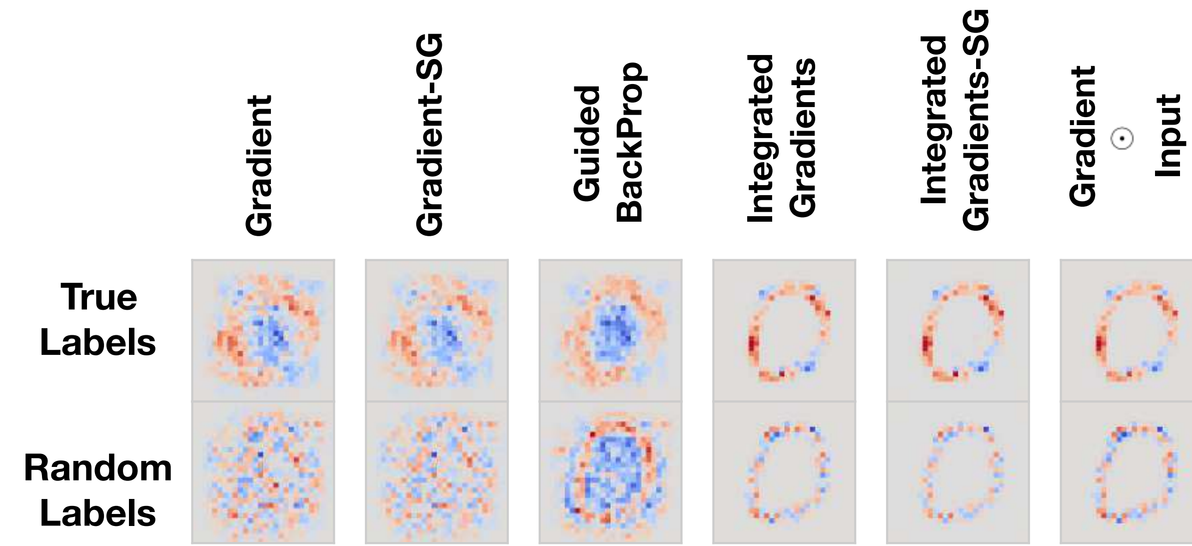
Data Randomization

MLP - MNIST

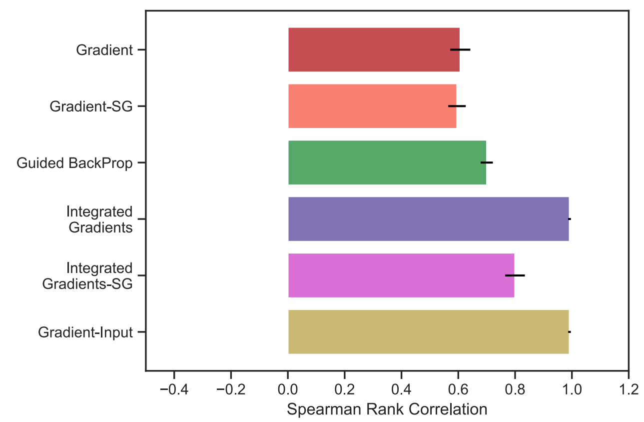
Absolute-Value Visualization



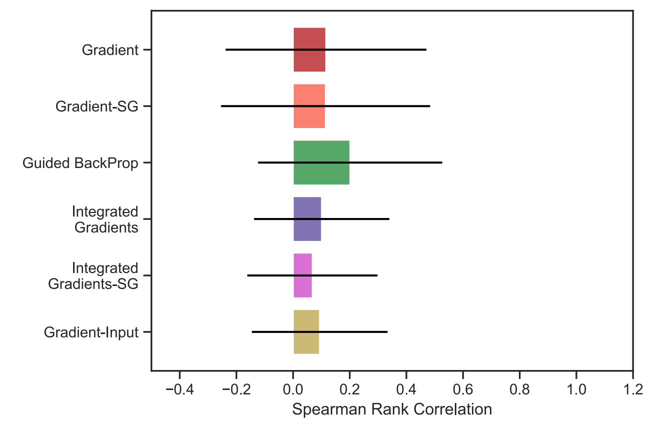
Diverging Visualization



Rank Correlation - Abs



Rank Correlation - No Abs



Some Insights

- Nie et. al. (ICML 2018) theoretical showed that Guided back propagation is doing input reconstruction.

- Observed in Mahendra et. al. 2014 (ECCV) as well.

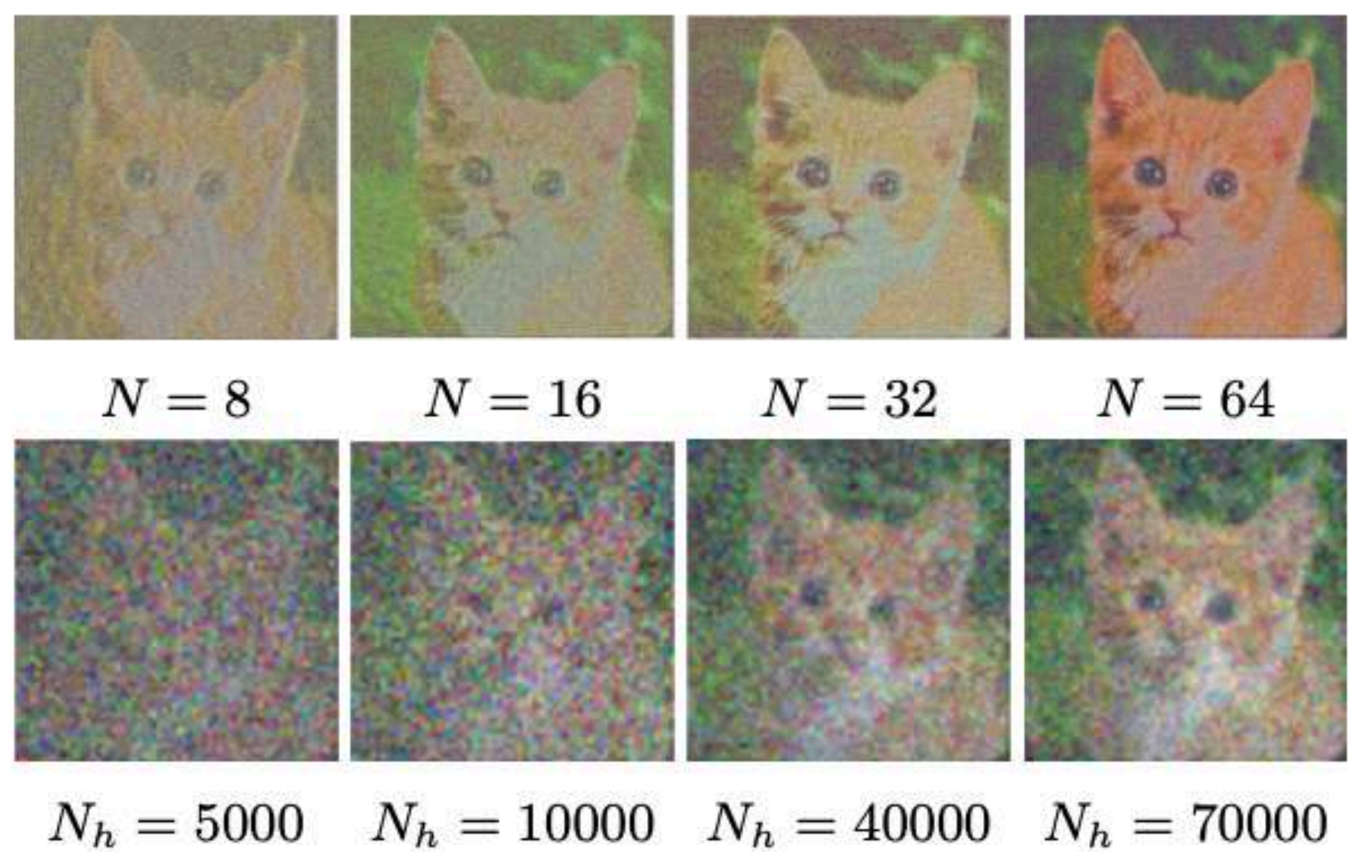
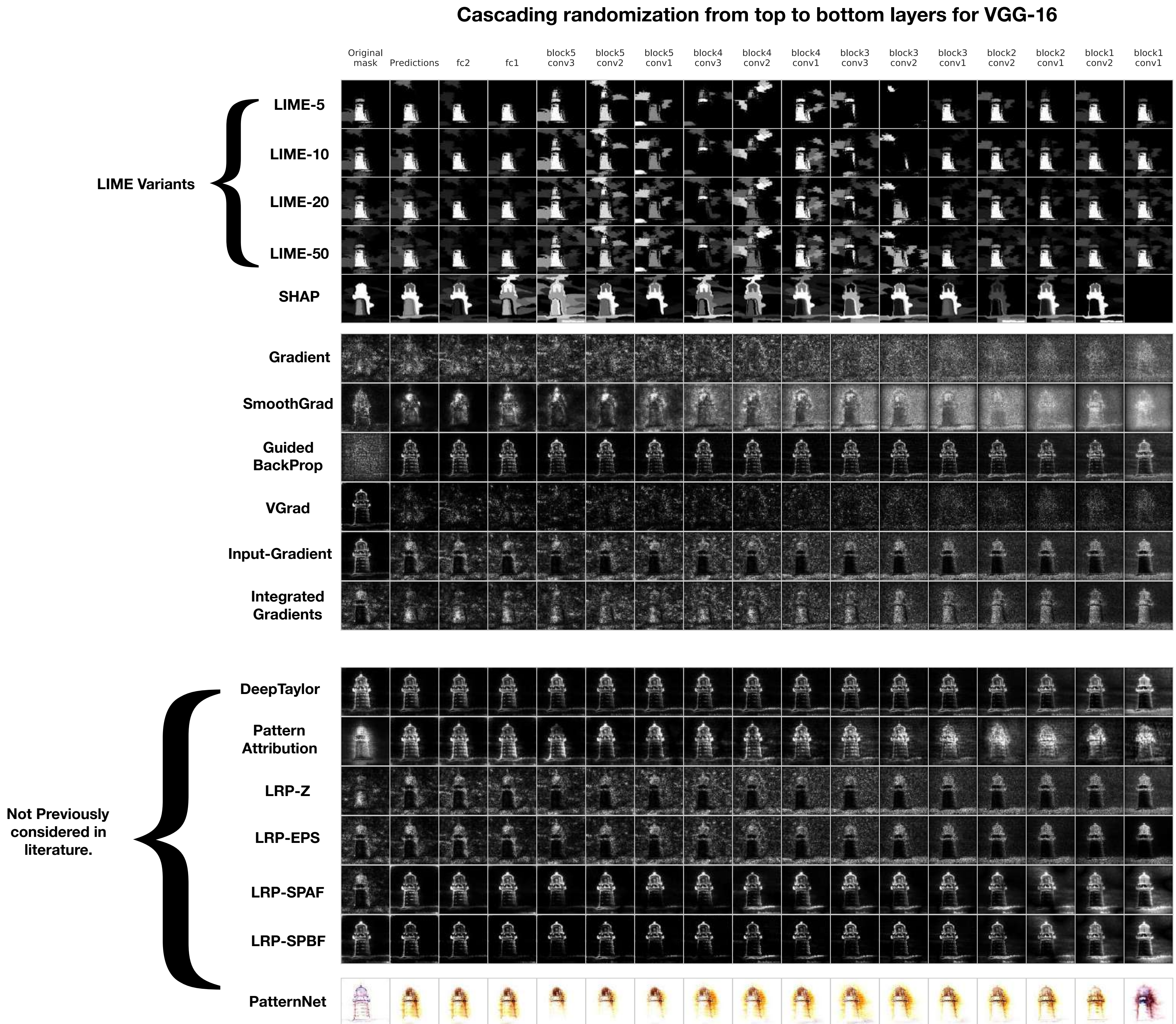


Figure from Nie et. al, 2018.

Summary

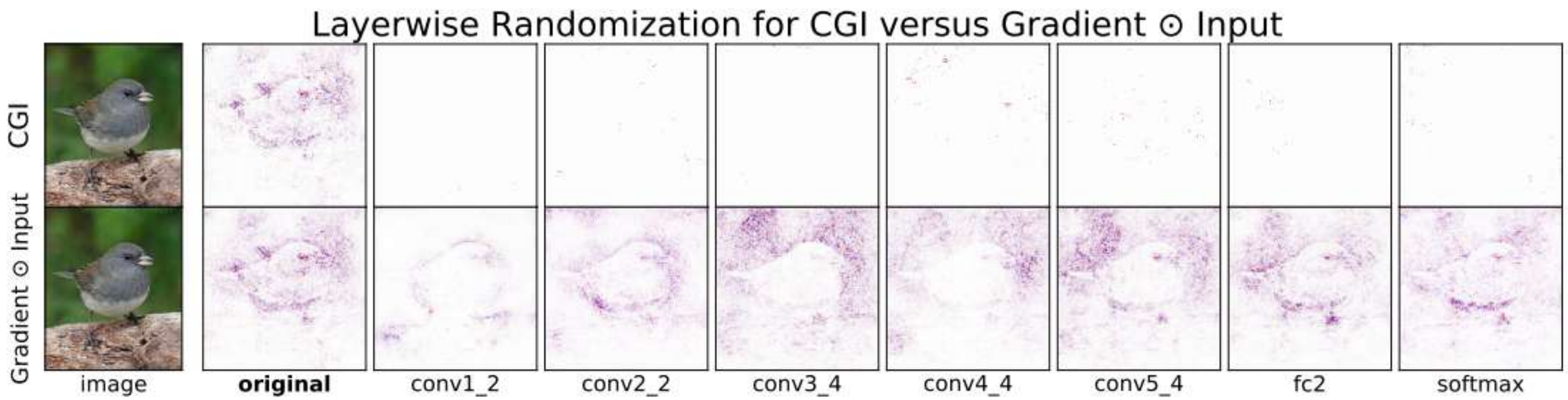
- We focused on gradient-based methods mostly.
- Sanity checks don't tell if a method is good, just if it is invariant.
- Sole visual inspection can be deceiving.

What about other methods



A Fix for Sanity Checks

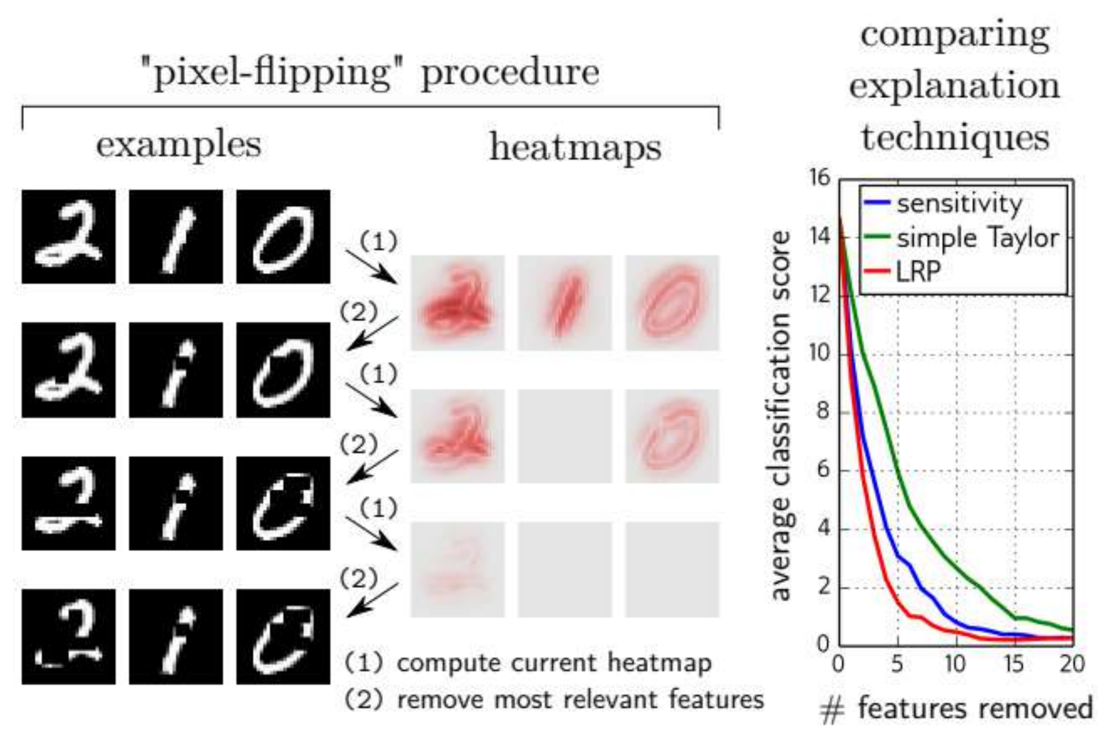
- **Gupta et. al.** fix this with competition for gradients (CGI).



[Figure from Gupta et. al. 2019.]

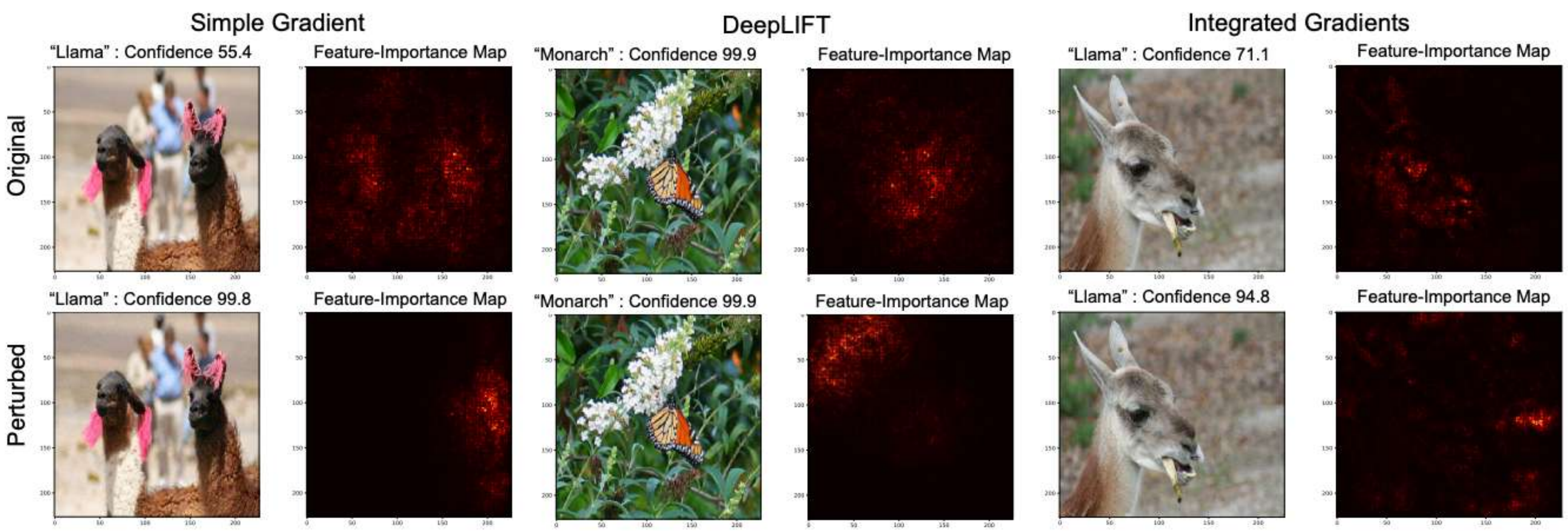
Other Assessment Methods

- Hooker et. al. (to appear at Neurips 2019) propose to remove and retrain.
- Adel et. al. propose FSM to 'quantify' information content.
- Yang et. al. introduce a benchmark (w/ground truth) and other metrics to assess how well a map captures model behavior.

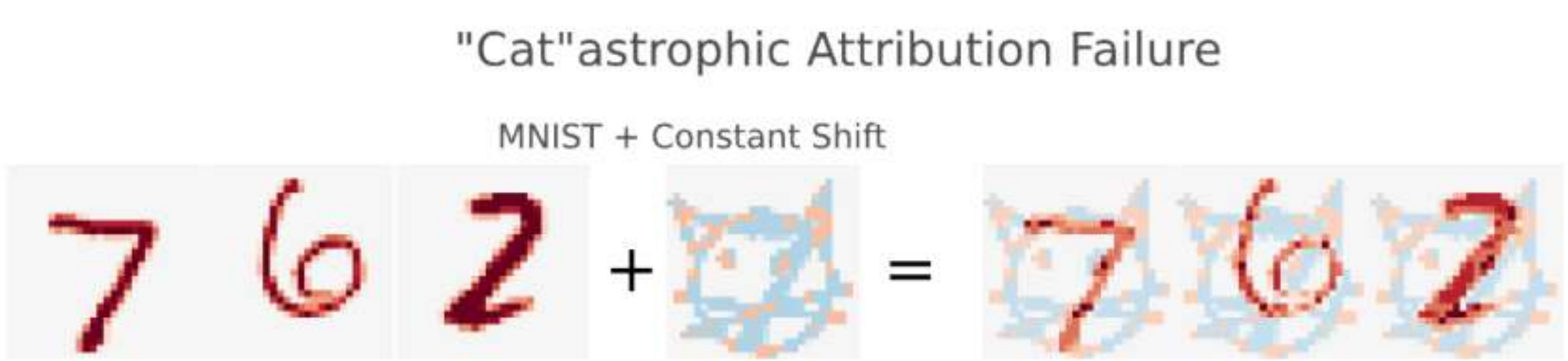


Attacks

- ‘Adversarial’ attack on explanations by Ghorbani et. al.



- Mean-shift attack by Kindermans & Hooker et. al.



Conundrum Persists

- For methods that pass sanity checks how do we choose among these?
- Can end-users (developers) use these methods to debug?
- What about other explanation classes (concepts and global methods)?