Advancements in Video Understanding

Interpretability & Attribution Methods

RESTRICTING THE FLOW:
INFORMATION BOTTLENECKS FOR ATTRIBUTION

Karl Schulz[1*†], **Leon Sixt**[2*], **Federico Tombari**[1], **Tim Landgraf**[2]
* contributed equally † work done at the Freie Universität Berlin
Technische Universität München[1] Freie Universität Berlin[2]
Corresponding authors: karl.schulz@tum.de, leon.sixt@fu-berlin.de

ABSTRACT

Attribution methods provide insights into the decision-making of machine learning models like artificial neural networks. For a given input sample, they assign a relevance score to each individual input variable, such as the pixels of an image. In this work, we adopt the information bottleneck concept for attribution. By adding noise to intermediate feature maps, we restrict the flow of information and can quantify (in bits) how much information image regions provide. We compare our method against ten baselines using three different metrics on VGG-16 and ResNet-50, and find that our methods outperform all baselines in five out of six settings. The method's information-theoretic foundation provides an absolute frame of reference for attribution values (bits) and a guarantee that regions scored close to zero are not necessary for the network's decision.

Proceedings of the International Conference on Learning Representations.

*(**ICLR**) 2020.*

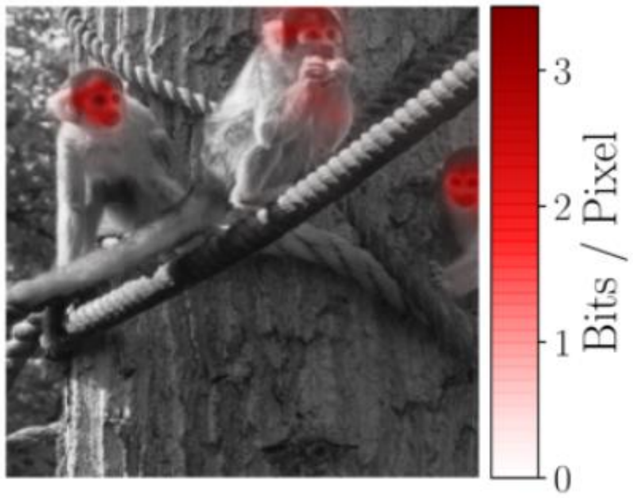
Paper: https://openreview.net/pdf?id=S1xWh1rYwB

Presenter: Giorgio Roffo

Date: 27/01/2022

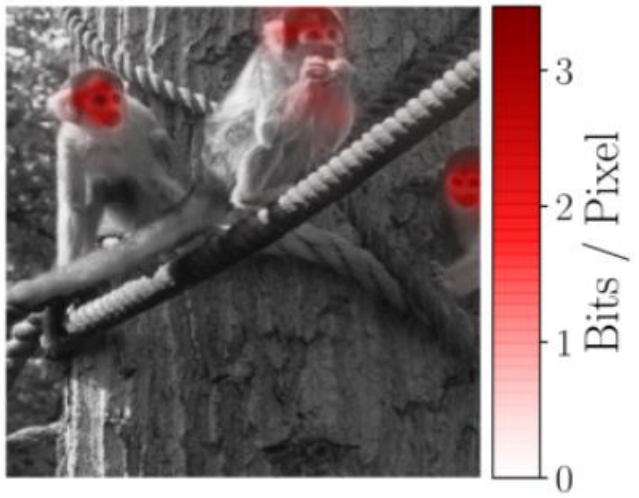
Attribution Methods

- Model interpretability is an important requirement (medical decision making or autonomous driving).
- **Attribution methods** (Selvaraju et al., 2017; Zeiler & Fergus, 2014; Smilkov et al., 2017) aim to **explain the model behavior** by assigning a relevance score to each input variable.
- Attribution methods **identify the pixels responsible** for the **classification** of the input image.
- The relevance scores can be visualized as heatmaps over the input.



Attribution Methods

- Model interpretability is an important requirement (medical decision making or autonomous driving).
- Attribution methods (Selvaraju et al., 2017; Zeiler & Fergus, 2014; Smilkov et al., 2017) aim to explain the model behavior by assigning a relevance score to each input variable.
- Attribution methods **identify the pixels responsible** for the **classification** of the input image.
- The relevance scores can be visualized as heatmaps over the input.



- For attribution, **no ground truth exists**.
- If an attribution heatmap highlights subjectively irrelevant areas:
  - Reflect the network's unexpected way of processing the data
  - Inaccurate heatmap

Related Work: Attribution Methods

- Several AMs in literature:

1. Gradient Maps
2. Saliency Maps

based on calculating the gradient of the target output neuron w.r.t. to the input features.

3. Smooth Grad
4. Integrated Grad.

5. Layerwise Relevance Propagation (LRP)
6. Deep Taylor Decomposition (DTD)
7. Guided Backpropagation (GuidedBP)
8. DeepLIFT

9. Pattern Attribution

10. Occlusion-14

11. Grad-CAM
12. Guided Grad-CAM
13. ...

Related Work: Attribution Methods

- Several AMs in literature:

1. Gradient Maps
2. Saliency Maps

based on calculating the gradient of the target output neuron w.r.t. to the input features.

3. Smooth Grad
4. Integrated Grad.

improve over gradient-based attribution maps by averaging the gradient of multiple inputs (e.g., a local neighborhood)

5. Layerwise Relevance Propagation (LRP)
6. Deep Taylor Decomposition (DTD)
7. Guided Backpropagation (GuidedBP)
8. DeepLIFT

9. Pattern Attribution

10. Occlusion-14

11. Grad-CAM
12. Guided Grad-CAM
13. ...

Related Work: Attribution Methods

- Several AMs in literature:

1. Gradient Maps
2. Saliency Maps

based on calculating the gradient of the target output neuron w.r.t. to the input features.

3. Smooth Grad
4. Integrated Grad.

improve over gradient-based attribution maps by averaging the gradient of multiple inputs.

5. Layerwise Relevance Propagation (LRP)
6. Deep Taylor Decomposition (DTD)
7. Guided Backpropagation (GuidedBP)
8. DeepLIFT

modify the propagation rule.

9. Pattern Attribution

10. Occlusion-14

11. Grad-CAM
12. Guided Grad-CAM
13. ...

Related Work: Attribution Methods

- Several AMs in literature:

1. Gradient Maps
2. Saliency Maps

based on calculating the gradient of the target output neuron w.r.t. to the input features.

3. Smooth Grad
4. Integrated Grad.

improve over gradient-based attribution maps by averaging the gradient of multiple inputs.

5. Layerwise Relevance Propagation (LRP)
6. Deep Taylor Decomposition (DTD)
7. Guided Backpropagation (GuidedBP)
8. DeepLIFT

modify the propagation rule.

9. Pattern Attribution

builds upon DTD by estimating the signal's direction for the backward propagation.

10. Occlusion-14

11. Grad-CAM
12. Guided Grad-CAM
13. ...

Related Work: Attribution Methods

- Several AMs in literature:

1. Gradient Maps
2. Saliency Maps

based on calculating the gradient of the target output neuron w.r.t. to the input features.

3. Smooth Grad
4. Integrated Grad.

improve over gradient-based attribution maps by averaging the gradient of multiple inputs.

5. Layerwise Relevance Propagation (LRP)
6. Deep Taylor Decomposition (DTD)
7. Guided Backpropagation (GuidedBP)
8. DeepLIFT

modify the propagation rule.

9. Pattern Attribution

builds upon DTD by estimating the signal's direction for the backward propagation.

10. Occlusion-14

Measures the importance as the drop in classification accuracy after replacing individual image patches with zeros.

11. Grad-CAM
12. Guided Grad-CAM
13. ...

Related Work: Attribution Methods

- Several AMs in literature:

1. Gradient Maps
2. Saliency Maps

based on calculating the gradient of the target output neuron w.r.t. to the input features.

3. Smooth Grad
4. Integrated Grad.

improve over gradient-based attribution maps by averaging the gradient of multiple inputs.

5. Layerwise Relevance Propagation (LRP)
6. Deep Taylor Decomposition (DTD)
7. Guided Backpropagation (GuidedBP)
8. DeepLIFT

modify the propagation rule.

9. Pattern Attribution

builds upon DTD by estimating the signal's direction for the backward propagation.

10. Occlusion-14

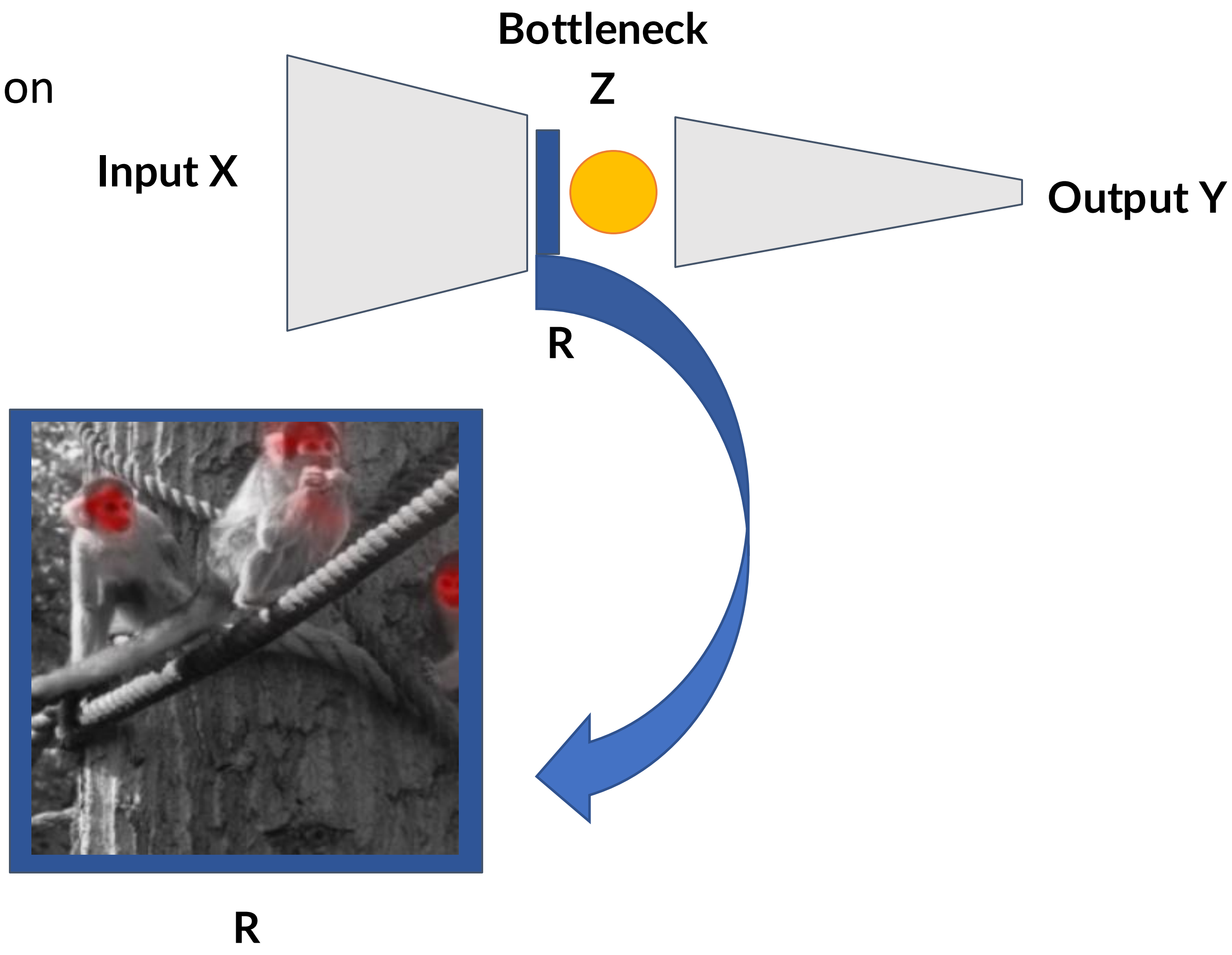
Measures the importance as the drop in classification accuracy after replacing individual image patches with zeros.

11. Grad-CAM
12. Guided Grad-CAM
13. ...

Take the activations of the final convolutional layer to compute relevance scores. Grad-Cam + GuidedBP = Guided Grad-CAM.

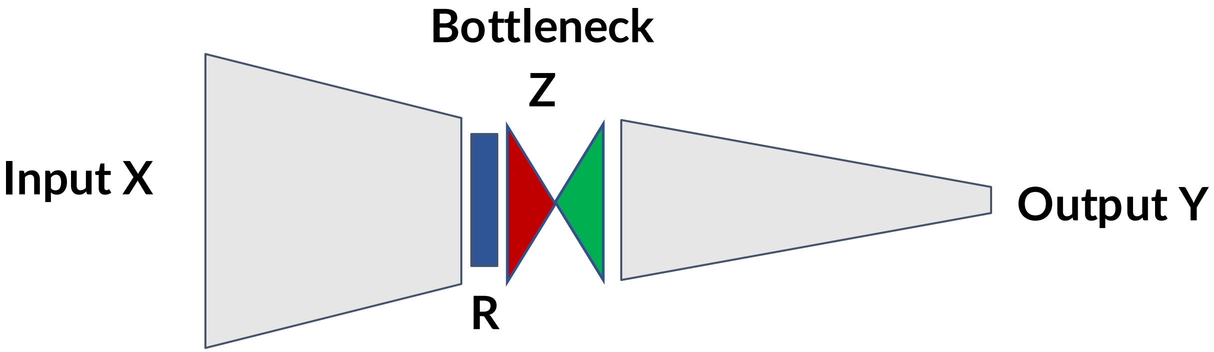
Main idea: Information Bottlenecks for Attribution (IBA)

- Given a pre-trained model **M**.
- It is possible to measure how much information image regions provide.



Main idea: Information Bottlenecks for Attribution (IBA)

- Given a pre-trained model **M**.
- It is possible to measure how much information image regions provide.
- Given a layer L in the network:
  - **IBA creates a bottleneck by injecting noise into the feature maps R.**
  - The intensity of the noise is optimized to minimize the information flow (bottleneck).
  - Simultaneously, the original objective is maximized.
- The parameters of the original model **M** are not changed.

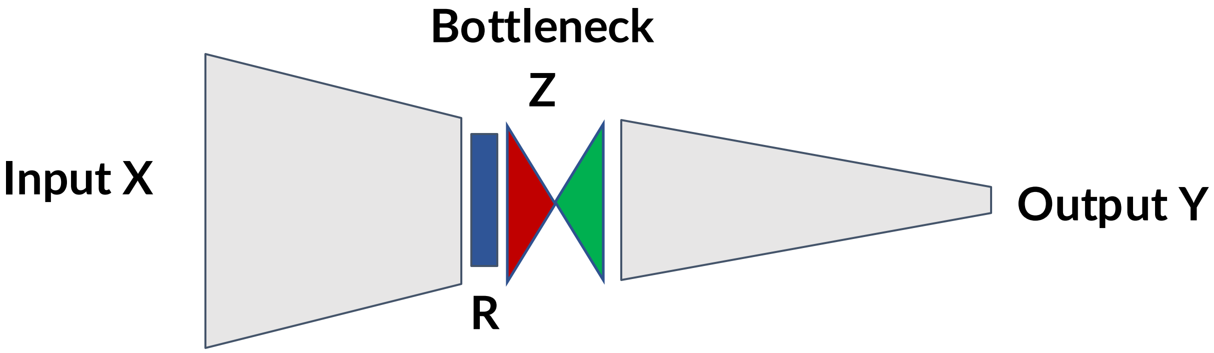


$$\text{MAX}\left[\ I(Z, Y; \boldsymbol{\theta}) - \alpha\, I(Z, X;\ \boldsymbol{\theta})\ \right]$$

Minimizes the amount of transmitted information while retaining a high classifier score for the explained class.

Main idea: Information Bottlenecks for Attribution (IBA)

- Given a layer L in the network:
  - **IBA creates a bottleneck by injecting noise into the feature maps R.**
  - The intensity of the noise is optimized to minimize the information flow (bottleneck).
  - Simultaneously, the original objective is maximized.
- **Example**:



$$\text{MAX}\left[\; I(Z, Y; \boldsymbol{\theta}) - \alpha\, I(Z, X; \boldsymbol{\theta}) \;\right]$$

Minimizes the amount of transmitted information while retaining a high classifier score for the explained class.

Information Bottlenecks for Attribution (IBA) 1/2

1. The information the new variable Z shares with the labels Y is **maximized I(Y,Z)**.
2. The information the variable Z shares with the labels X is **minimized I(X,Z)**.
3. For the ResNet the bottleneck is added after conv3_* layer.
4. Let **R** denote the intermediate representations at the L-th layer. $\mathbf{R = f_L(X)}$ where f_L is the L-th layer output.
5. When increasing the noise, the signal R is partly replaced with noise.
6. Noise $\mathbf{\mathcal{N} = \mathcal{N}(\mu_R,\sigma_R)}$ where μ_R,σ_R estimated for any Rs empirically

$$\mathbf{Z} = \lambda(X)\mathbf{R} + (1 - \lambda(X))\,\mathcal{N}$$

- **Where** λ(X) controls the damping of the signal and the addition of the noise ($\lambda_i \in [0, 1]$).

Information Bottlenecks for Attribution (IBA) 2/2

$$\mathbf{Z} = \boldsymbol{\lambda(X)}\mathbf{R} + (1 - \boldsymbol{\lambda(X)})\, \mathcal{N}$$

- To estimate the information Z still contains about R mutual information I[R, Z] is used:

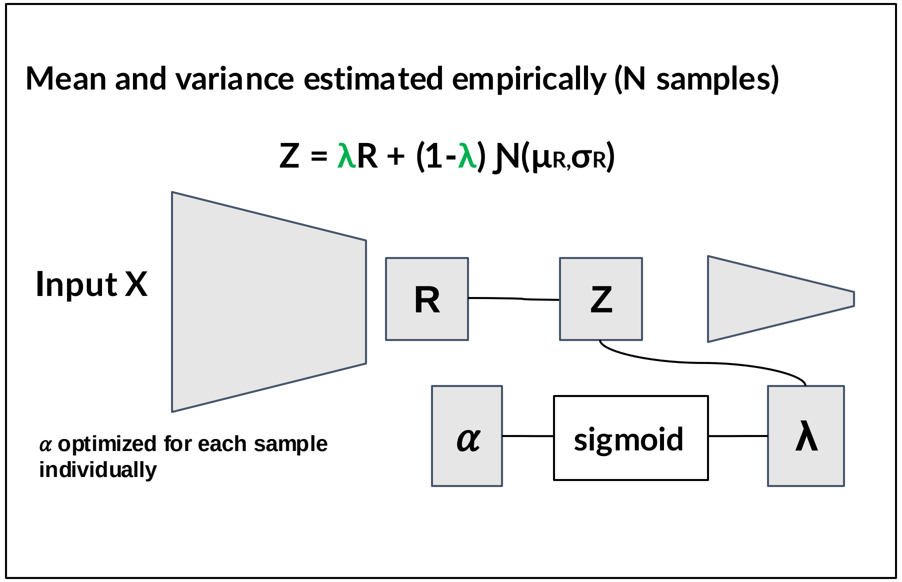
$$I[R, Z] = E_R[D_{KL}[\, P(Z|R) \,||\, Q(Z) \,]]$$

- P(Z|R) this is R with noise.
- $Q(Z) = \mathcal{N}(\mu_R, \sigma_R)$ this is pure noise.
- The information loss function is therefore:

$$\mathbf{L_{INF} = E_R[D_{KL}[\, P(Z|R) \,||\, Q(Z) \,]]}$$

- Then, we obtain the following optimization problem:

$$\mathbf{L = L_{CE} + \beta L_{INF}}$$



Minimizes the amount of transmitted information while retaining a high classifier score for the explained class.

*** IBA quantifies (in bits) how much information image regions provide.**

Two IBA Strategies

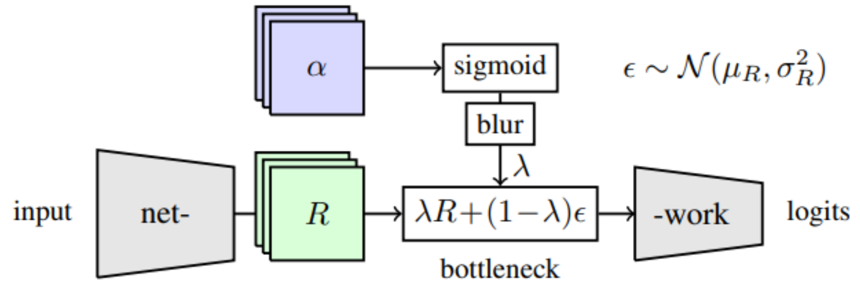


Figure 2: *Per-Sample Bottleneck*: The mask (blue) contains an α_i for each r_i in the intermediate feature maps R (green). The parameter α controls how much information is passed to the next layer. The mask α is optimized for each sample individually according to equation 6.

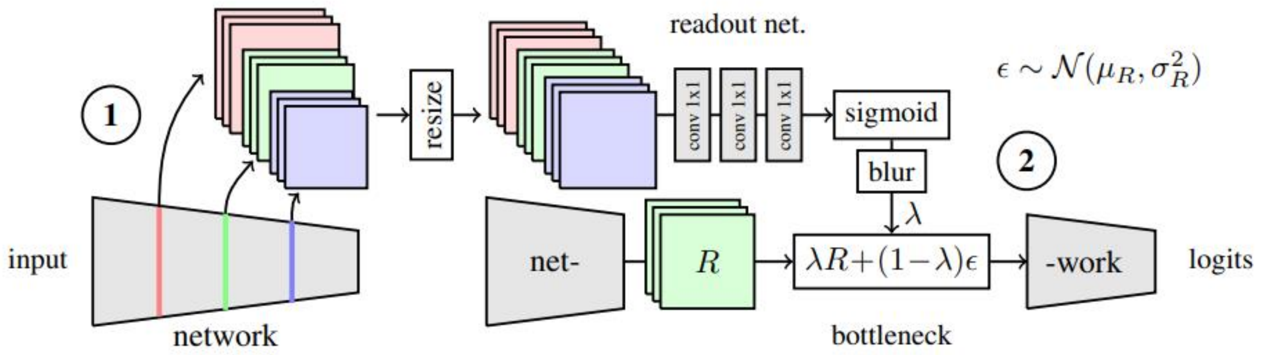
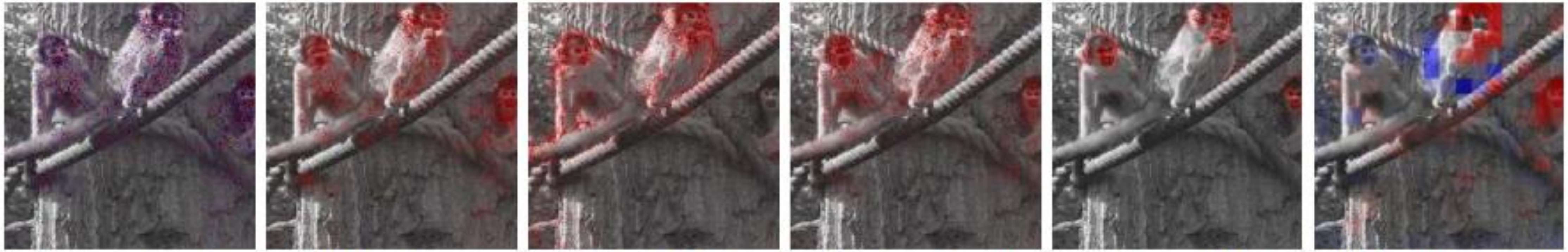


Figure 3: *Readout Bottleneck*: In the first forward pass (1), feature maps are collected at different depths. The readout network uses a resized version of the feature maps to predict the parameters for the bottleneck layer. In the second forward pass (2), the bottleneck is inserted and noise added. All parameters of the analyzed network are kept fixed.

Experiments: QUALITATIVE ASSESSMENT



(a) Gradient (b) Saliency (c) SmoothGrad (d) Int. Grad. (e) GuidedBP (f) Occlusion-14

(g) Grad-CAM (h) G.Grad-CAM (i) PatternAttr. (j) LRP (k) *Per-Sample* (l) *Readout*

Figure 5: Heatmaps of all implemented methods for the VGG-16 (see Appendix A for more).

Experiments: QUALITATIVE ASSESSMENT

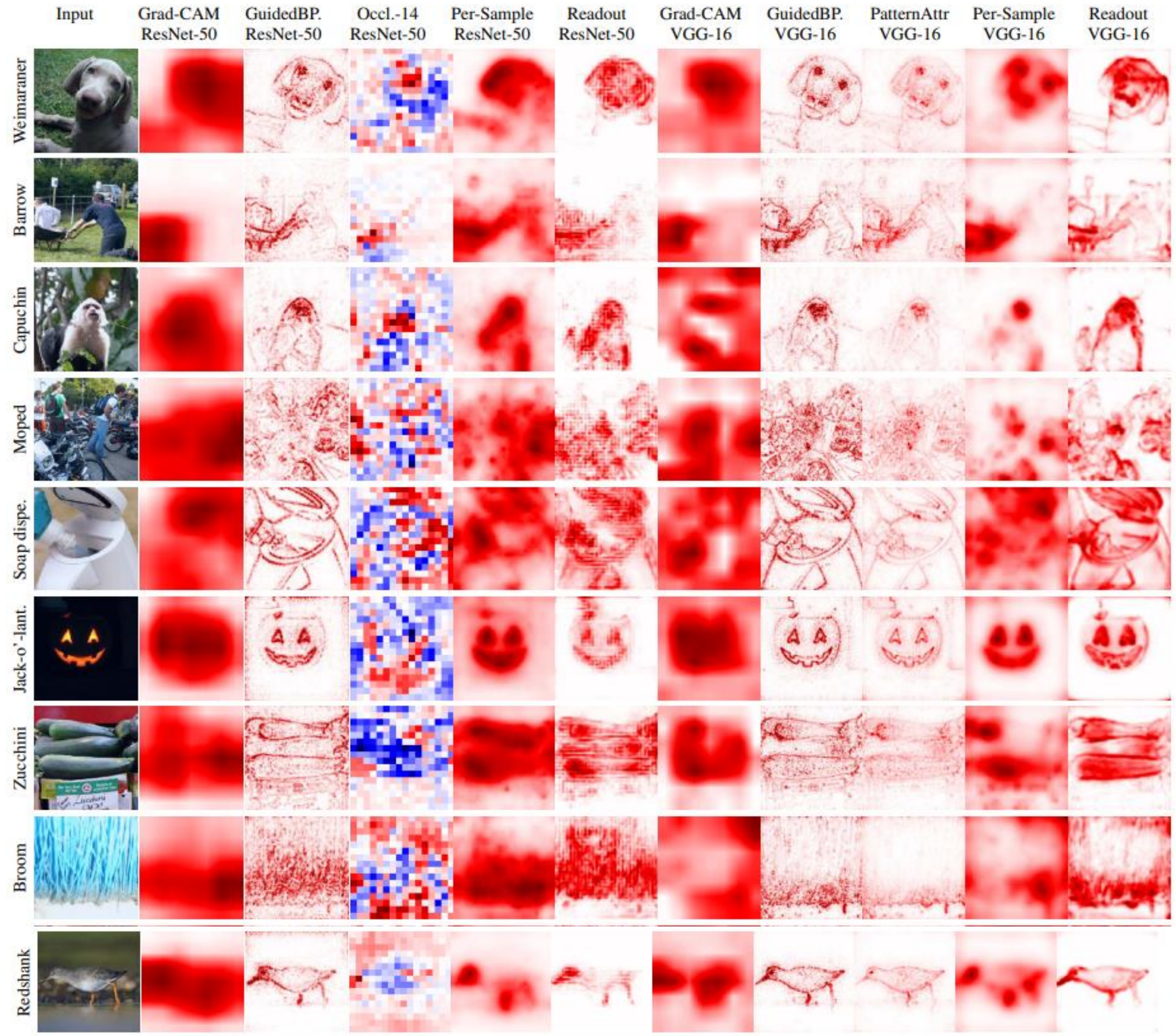
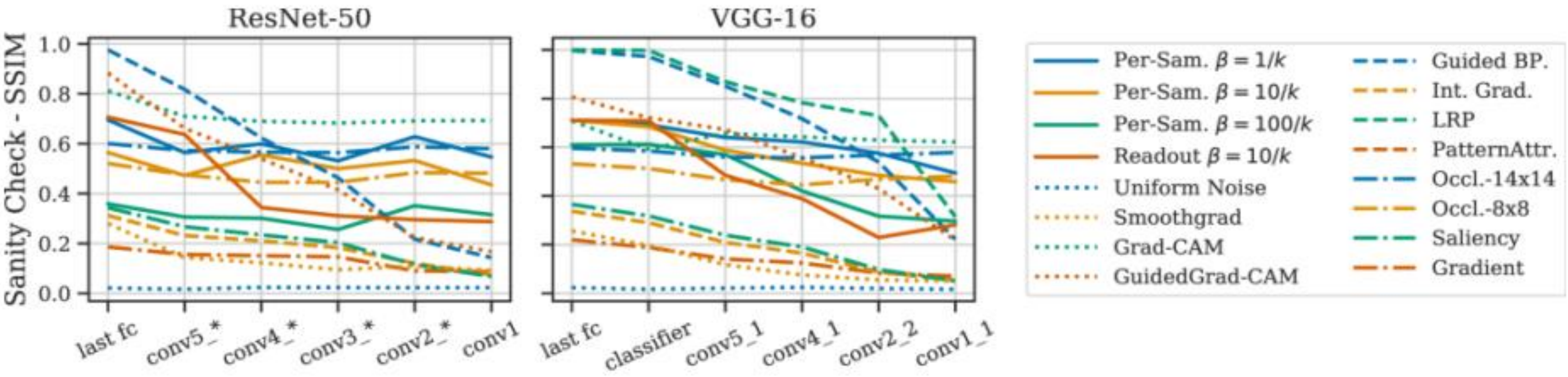


Figure 8: Blue indicates negative relevance and red positive. The authors promise that the samples were picked truly randomly, no cherry-picking, no lets-sample-again-does-not-look-nice-enough.

So they don't use the standard sampling strategy LSADNLNE

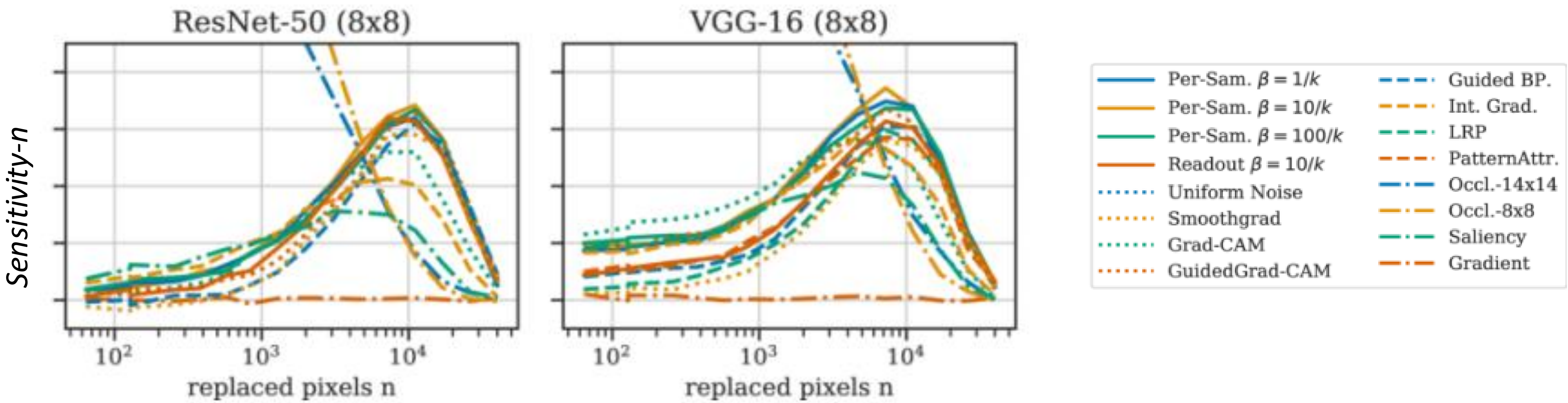
Experiments: RANDOMIZATION OF MODEL PARAMETERS

- A sound attribution method should depend on the entire network's parameter set, Adebayo et al. (2018).
- Starting from the last layer, an increasing proportion of the network parameters is re-initialized until all parameters are random.
- The difference between the original heatmap and the heatmap obtained from the randomized model is quantified using SSIM.



Experiments: SENSITIVITY-N

- Sensitivity-n masks the network's input randomly and then measures how strongly the amount of attribution in the mask correlates with the drop in classifier score.
- Given a set T_n containing n randomly selected pixel indices (pixels T_n = 0), Sensitivity-n measures the Pearson correlation coefficient between
  - The relevance at pixel i (given by the attribution method)
  - The difference between the classifier logit output for class c S[x] and S[x with n zero pixels].
- Per-Sample Bottlenecks perform best for both models when more then 2% of all pixels are masked.



Experiments: IMAGE DEGRADATION and BOUNDING BOX

- **Image Degradation:**
  - Given an attribution heatmap, the input is split in tiles, which are ranked by the sum of attribution values within each corresponding tile of the attribution.
  - At each iteration, the highest-ranked tile is replaced with a constant value, the modified input is fed through the network, and the resulting drop in target class score is measured.
  - The score is then normalized between [0, 1].

- **Bounding Box:**
  - To quantify how well attribution methods identify and localize the object of interest.
  - If the bounding box contains n pixels, we measure how many of the n-th highest scored pixels are contained in the bounding box. By dividing by n, we obtain a ratio between 0 and 1.

Experiments: IMAGE DEGRADATION (ResNet-50)

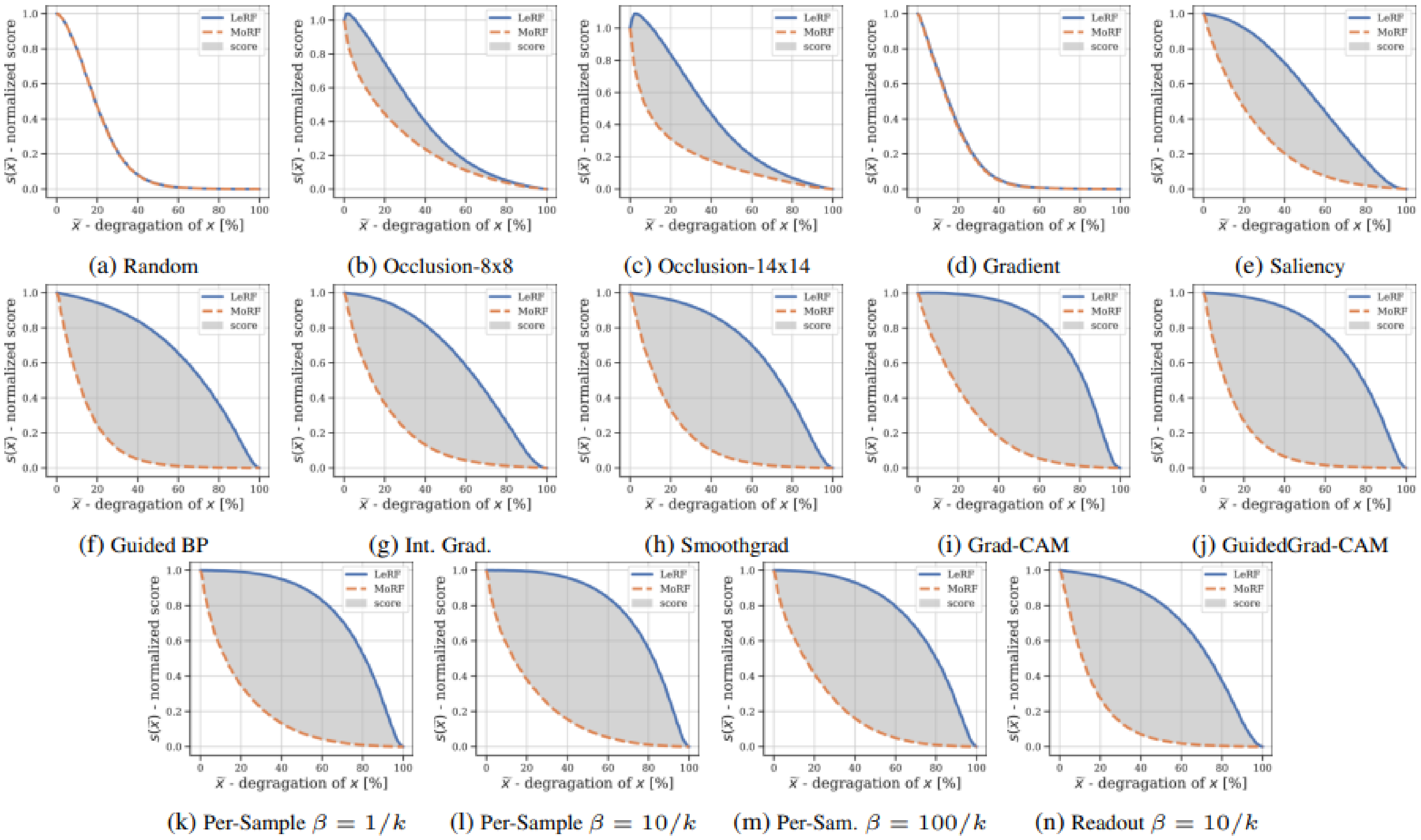


Figure 11: MoRF and LeRF for the ResNet-50 network using 14x14 tiles.

Experiments: IMAGE DEGRADATION (VGG-16)

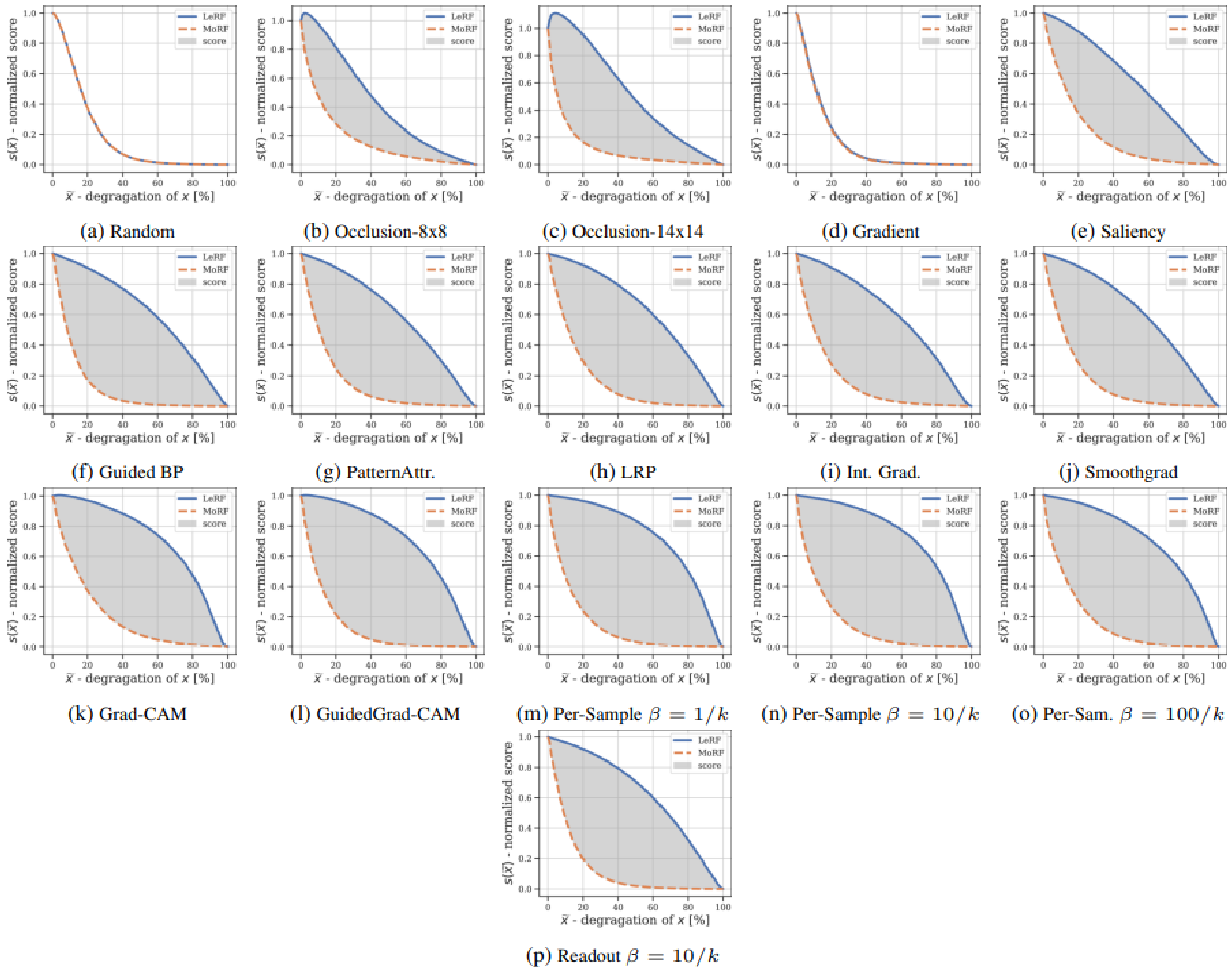


Figure 12: MoRF and LeRF paths for the VGG-16 network using 14x14 tiles.

Experiments: BOUNDING BOX and IMAGE DEGRADATION

Model & Evaluation	ResNet-50 deg.		VGG-16 deg.		ResNet	VGG
	8x8	14x14	8x8	14x14	bbox	bbox
Random	0.000	0.000	0.000	0.000	0.167	0.167
Occlusion-8x8	0.162	0.130	0.267	0.258	0.296	0.312
Occlusion-14x14	0.228	0.231	0.402	0.404	0.341	0.358
Gradient	0.002	0.005	0.001	0.005	0.259	0.276
Saliency	0.287	0.305	0.326	0.362	0.363	0.393
GuidedBP	0.491	0.515	0.460	0.493	0.388	0.373
PatternAttribution	–	–	0.440	0.457	–	0.404
LRP $\alpha=1, \beta=0$	–	–	0.471	0.486	–	0.397
LRP $\alpha=0, \beta=1, \epsilon=5$	–	–	0.462	0.467	–	0.441
Int. Grad.	0.401	0.424	0.420	0.453	0.372	0.396
SmoothGrad	0.485	0.502	0.438	0.455	0.439	0.399
Grad-CAM	0.536	0.541	0.510	0.517	0.465	0.399
GuidedGrad-CAM	0.565	**0.577**	0.555	0.576	0.468	0.419
IBA Per-Sample $\beta=1/k$	**0.573**	0.573	0.581	0.583	0.606	0.566
IBA Per-Sample $\beta=10/k$	0.572	0.571	**0.582**	**0.585**	**0.620**	**0.593**
IBA Per-Sample $\beta=100/k$	0.534	0.535	0.542	0.545	0.574	0.568
IBA Readout $\beta=10/k$	0.536	0.536	0.490	0.536	0.484	0.437

Table 1: *Degradation (deg.)*: Integral between LeRF and MoRF in the degradation benchmark for different models and window sizes over the ImageNet test set. *Bounding Box (bbox)*: the ratio of the highest scored pixels within the bounding box. For ResNet-50, we show no results for PatternAttribution and LRP as no PyTorch implementation supports skip-connections.

Conclusions

- Model interpretability is an important requirement (medical decision making or autonomous driving).
- Attribution methods (Selvaraju et al., 2017; Zeiler & Fergus, 2014; Smilkov et al., 2017) aim to explain the model behavior by assigning a relevance score to each input variable.
- AIB Identifies the pixels responsible for the classification of the input image.
- A **bottleneck layer** is used to **inject noise** into a given feature layer.
- **Minimizes** the amount of **transmitted information** while **retaining** a **high classifier score** for the explained class (**MAX [$I(Z, Y; \theta) - \alpha\, I(Z, X; \theta)$]**)
- AIB can quantify (in **bits**) how much information image regions provide.
- The Per-Sample Bottleneck is optimized per single data point, whereas the Readout Bottleneck is trained on the entire dataset.

Thank you

Any questions?