Supplementary Material for Occlusion Sensitivity Analysis with Augmentation Subspace Perturbation in Deep Feature Space

1. Theory

In this section, we present details of our method with increased mathematical formality. Some of the actual discussion and contributions may be repeated to complement the main idea.

1.1. Actual causality

Actual causality [6] is a framework to formally explain the way model predictions depend on input variables, what are the output causes, and how certain changes in the inputs can change the predictions. It extends counterfactual reasoning with contingencies, which means that if a Boolean function $\phi\left(x_1, x_2, \cdots, x_n\right)$ changes when a variable x_i is altered, then ϕ depends on x_i.

Moreover, the Degree of Responsibility r is a quantification of causality [1], which is based on the size k of the smallest contingency required to create a counterfactual dependency [2], *i.e.*, the minimal change to alter the function output.

Definition 1.1 (Singleton cause)**.** Let f be a machine learning model and x an input, an entry $x_{i_1\, i_2\, \cdots\, i_n}$ (think a pixel) is a cause of $f\left(x\right)$ *if and only if* there is a subset $\chi \subset x$ such that the following hold [2, p. 3]:

1. $x_{i_1\, i_2\, \cdots\, i_n} \notin \chi$
2. output invariance to masking of χ:
$$\text{Let } \chi' \subset \chi, m \in \mathbb{R}, \text{ then } \chi' = m \implies \Delta f = 0$$
3. output dependency to masking of $x_{i_1\, i_2\, \cdots\, i_n}$:
$$\text{Let } m \in \mathbb{R}, \text{ then } \chi = x_{i_1\, i_2\, \cdots\, i_n} = m \implies \Delta f \neq 0$$

Definition 1.2 (Cause witness)**.** If a subset $\chi \in x$ and entry $x_{i_1\, i_2\, \cdots\, i_n}$ satisfy Def. 1.1, then we say χ is a witness to the fact $x_{i_1\, i_2\, \cdots\, i_n}$ is a cause of x [2, p. 3].

Definition 1.3 (Simplified Degree of Responsibility)**.** If a subset $\chi \in x$ and entry $x_{i_1\, i_2\, \cdots\, i_n}$ satisfy the definition of Singleton Cause from [2], then

$$r\left(x_{i_1\, i_2\, \cdots\, i_n} | x, f\right) = \frac{1}{1+k}, \tag{1}$$

where k is the size of the minimal witness, which refers to the smallest subset of input variables that, when changed, can demonstrate that a particular input variable has an effect on the output of a function.

In our terminology, interpreters are algorithms which process a model and a single input, and output an explanation. In computer vision, a valid explanation could be an attribution heatmap over the original input image. Next, we formally define this concept.

Definition 1.4 (Explanation)**.** Let f be a machine learning model and $\mathbf{x}$ an input with output $f\left(\mathbf{x}\right)$, and $\mathbf{S} = \mathbf{x} \odot \mathbf{M}$ a masked subset of the input. Then, the explanation $\mathbf{E}$ of the model f given input $\mathbf{x}$ is the minimal subset which maintains the output [2]

$$\mathbf{E}\left(f | \mathbf{x}\right) = \min_{|\mathbf{S}|} \mathbf{S} : f\left(\mathbf{S}\right) = f\left(\mathbf{x}\right), |\mathbf{S}| > 0 \tag{2}$$

where $|\cdot|$ is the number of items in the set, *e.g.*, the number of non-masked pixels

Remark 1 (Triviality). The explanation must not be a null tensor. Any model will output a prediction for the null tensor, however this would be a trivial explanation for all inputs sharing the same prediction. For example, if a model outputs "dog" for the null tensor, then all images of dogs would have an empty heatmap as explanation.

Remark 2 (Non-uniqueness). The explanation may not be unique. Inputs might have symmetries or repetitions, leading to multiple viable subsets of the same size.

However, computing an explanation is NP-complete and real interpreters will output approximate explanations [2].

Definition 1.5 (Approximate Explanation)**.** Let f be a machine learning model and $\mathbf{x}$ an input with output $f\left(\mathbf{x}\right)$. The approximate explanation $\tilde{\mathbf{E}}$ of model f given input $\mathbf{x}$ is the probability distribution indicating if the input entry belongs to the explanation $\mathbf{x}$.

$$\tilde{\mathbf{E}}_{i_1\, i_2\, \cdots\, i_n} = p\left(\mathbf{x}_{i_1\, i_2\, \cdots\, i_n} \in \mathbf{x}\right) \tag{3}$$

Remark 3 (Normalization).

$$\sum_{i_1\, i_2\, \cdots\, i_n} \tilde{\mathbf{E}}_{i_1\, i_2\, \cdots\, i_n} = 1 \tag{4}$$

Remark 4 (Approximate Explanation and Degree of responsibility). While explanations can be seen as a binary inclusion mask, *i.e.*, the mask is 1 if the entry belongs to the explanation. However, the degree of responsibility is a measure between 0 and 1, so it can be seen as a proxy for probability.

Remark 5 (Composition). Given the probabilistic nature of an approximate explanation, the composed approximate explanation can be built when multiple approximate explanations are available (*e.g.*, when multiple interpreters can be used). The composed explanation can be built by simple summation and renormalization following Rem. 3.

Definition 1.6 (Minimal Size)**.** Let $\tilde{\mathbf{E}}$ be the approximate explanation of model f given input $\mathbf{x}$. The explanation minimal size is defined as

$$s_{min}\left(\tilde{\mathbf{E}}\left(f|\mathbf{x}\right)\right) = \frac{|\mathbf{S}|}{|\mathbf{x}|}, \text{ with } f\left(\mathbf{S}\right) \approx f\left(\mathbf{x}\right) \tag{5}$$

Remark 6 (Minimal Size metric). Def. 1.6 is a viable explanation metric. It can be measured by Algorithm 1. In that sense, an explanation with low minimal size indicates the most substantial region for the model was reached.

Algorithm 1 Minimal Size Metric Computation

```
Require: $\mathbf{x} \leftarrow$ image, $f \leftarrow$ model, $\mathbf{H} \leftarrow$ heatmap, $s \leftarrow$ number of steps, $\delta \leftarrow$ tolerance
  for $i \leftarrow 1$ to $|\mathbf{x}|$ in $s$ steps do
      $\mathbf{S} \leftarrow$ top $i$ pixels from $\mathbf{x}$ based on $\mathbf{H}$
      if $||f\left(\mathbf{x}\right) - f\left(\mathbf{S}\right)||_1 \leq \delta$ then return $|\mathbf{S}|/|\mathbf{x}|$
      end if
  end for
```

1.2. Occlusion Sensitivity Analysis

Occlusion computes explanation heatmaps by replacing image regions with a given baseline (masking it to 0), and measuring the score difference in the output [10, 14].

Proposition 1 (Occlusion degree of responsibility)**.** Let f be a model which outputs a probability score $p \in [0, 1]$, $\mathbf{x}$ an input and $\mathbf{M}$ as binary mask with the same shape as $\mathbf{x}$, and $\odot$ be the Hadamard product. Then, the degree of responsibility of the masked region is

$$r = 1 - \frac{p\left(\mathbf{x} \odot \mathbf{M}\right)}{p\left(\mathbf{x}\right)}$$

Proof. Let the degree of responsibility be $r = \frac{1}{1+k}$ (Def. 1.3), the factor k represents the size of the minimal witness [2], which should be 0 for relevant causes and ∞ for irrelevant ones, *i.e.*, $k \in [0, +\infty)$.

First, assume that for every image, there is a defined region $\mathbf{S}$ where it's minimal witness has size 0, *i.e.*, when we mask all of the image keeping nothing but $\mathbf{S}$, the prediction score p output by the model is unaltered. Moreover, assume that the opposite action is also true: masking $\mathbf{S}$ while keeping any other part of the image will lead to a prediction collapse. Simply put,

$$p\left(\mathbf{x} - \mathbf{S}\right) = 0 \iff p\left(\mathbf{S}\right) = p\left(\mathbf{x}\right)$$

Conversely, if $\mathbf{S}'$ is an irrelevant area, masking it should render no change,

$$p\left(\mathbf{x} - \mathbf{S}'\right) = p\left(\mathbf{x}\right) \iff p\left(\mathbf{S}'\right) = 0$$

From such assumption, we should expect that the masked region minimal witness [2] size must be proportional to the score $p\left(\mathbf{S}\right)$ of keeping only the cause $\mathbf{S}$, and inversely proportional to the score $p\left(\mathbf{S}'\right)$ of keeping only an unimportant region $\mathbf{S}'$.

Thus, we can say

$$k \propto \frac{p\left(\mathbf{S}\right)}{p\left(\mathbf{S}'\right)} \equiv \frac{p\left(\mathbf{x} - \mathbf{S}\right)}{|p\left(\mathbf{x}\right) - p\left(\mathbf{x} - \mathbf{S}'\right)|} \equiv \frac{p\left(\mathbf{x} \odot \mathbf{M}\right)}{|p - p\left(\mathbf{x} \odot \mathbf{M}\right)|},$$

which should be 0 when the cause is masked and diverge when masking an unimportant region (score does not change).

Finally, we can effectively ignore the modulo assuming $p\left(\mathbf{x}\right) \geq p\left(\mathbf{x} \odot \mathbf{M}\right)$, and

$$\begin{aligned} r &= \frac{1}{1+k} \\ &= \frac{1}{1 + \frac{p(\mathbf{x} \odot \mathbf{M})}{p - p(\mathbf{x} \odot \mathbf{M})}} \\ &= \frac{p\left(\mathbf{x}\right) - p\left(\mathbf{x} \odot \mathbf{M}\right)}{p\left(\mathbf{x}\right)} \\ &= 1 - \frac{p\left(\mathbf{x} \odot \mathbf{M}\right)}{p\left(\mathbf{x}\right)} \end{aligned} \tag{6}$$

■

Proposition 2 (Occlusion approximate explanation)**.** The approximate explanation of an input for a single mask is

$$\tilde{\mathbf{E}}^{\mathbf{M}}\left(f|\mathbf{x}\right) = \left(1 - \mathbf{M}\right)\left(1 - \frac{p\left(\mathbf{x} \odot \mathbf{M}\right)}{p\left(\mathbf{x}\right)}\right)$$

Proof. Given $\mathbf{M}$ is a binary mask, $1 - \mathbf{M}$ is the inverse mask, *i.e.*, the masked region is set to 1. Then, from Prop. 1 and Rem. 4 we say the probability of the cause belonging to the masked region is equal the term $1 - \frac{p(\mathbf{x} \odot \mathbf{M})}{p(\mathbf{x})}$. ■

Lemma 1.1 (Occlusion Sensitivity Analysis (OSA))**.** The non-normalized general occlusion sensitivity analysis is the combination of individual occlusion explanations.

$$\tilde{\mathbf{E}}\left(f|\mathbf{x}\right) = \sum_i \left(1 - \mathbf{M}_i\right)\left(1 - \frac{p\left(\mathbf{x} \odot \mathbf{M}_i\right)}{p\left(\mathbf{x}\right)}\right) \tag{7}$$

Proof. Direct from Prop. 2 and Rem. 5 ■

OSA is a simple example of a perturbation-based method, in which the explanation is a composition of output scores relative variations for each masked input. Prop. 1 defines a way of computing the responsibility of a singular mask, *i.e.*, the probability it belongs to the cause, and Algorithm 2 shows how to compose it into an approximate explanation. OSA pseudocode can be found in the supplementary material.

Algorithm 2 Occlusion Sensitivity Analysis

```
Require: $\mathbf{x} \leftarrow$ image, $f \leftarrow$ model
  $n \leftarrow$ number of masks
  $l \leftarrow$ mask size
  $p \leftarrow f(\mathbf{x})$
  $\mathbf{H} \leftarrow 0$
  for $i \leftarrow 1$ to $n$ do
      $\mathbf{M} \leftarrow \text{mask}(i, \mathbf{x}.shape, l)$        ▷ any mask generator
      $\mathbf{x}^{\mathbf{M}} \leftarrow \mathbf{x} \odot \mathbf{M}$
      $p^{\mathbf{M}} \leftarrow f(\mathbf{x}^{\mathbf{M}})$
      $\mathbf{H} \leftarrow \mathbf{H} + (1 - \mathbf{M})\left(1 - \frac{p^{\mathbf{M}}}{p}\right)$        ▷ compose (Rem. 5)
  end for
return $\frac{\mathbf{H}}{\sum \mathbf{H}}$        ▷ normalize explanation
```

However, notice the formulation at Lem. 1.1 is different to the more traditional one defined by [10] in equation 6. The difference is mostly due to the different assumptions we took, but they can be shown to be proportionally equivalent, differing only by a constant. However, the formulation at Lem. 1.1 will be important for the methods we will propose next.

1.3. Occlusion Sensitivity Analysis with Deep Feature Vectors

Lem. 1.1 is a class-specific algorithm. However, most machine learning models actually output general vectors, also known as deep feature vector, which encode the input through the model. These vectors are then later processed to obtain the probability score of a single feature (class).

Proposition 3 (Representation degree of responsibility)**.** Let f be a model which outputs a vector $f(\mathbf{x}) = \mathbf{v} = (v_i), i \in [m]$. The degree of responsibility of the masked region is

$$r = \frac{||\mathbf{v} - \mathbf{v}(\mathbf{x} \odot \mathbf{M})||_p}{||\mathbf{v}||_p},$$

where $||\mathbf{v}||_p$ stands for the ℓ_p-norm

Proof. Let f be a model which outputs a scalar probability score $p(\mathbf{x}) \in [0, 1]$. Then, from Prop. 1, the degree of responsibility is

$$\begin{aligned} r &= 1 - \frac{p(\mathbf{x} \odot \mathbf{M})}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}) - p(\mathbf{x} \odot \mathbf{M})}{p(\mathbf{x})} \\ &= \frac{f(\mathbf{x}) - f(\mathbf{x} \odot \mathbf{M})}{f(\mathbf{x})} \end{aligned} \tag{8}$$

Now, notice that Prop. 1 assumes $f(\mathbf{x} \odot \mathbf{M}) \leq f(\mathbf{x})$ and $p(\mathbf{x}) \geq 0$. Then, we can extend this concept to a new f which outputs vectors by

$$\begin{aligned} \frac{f(\mathbf{x}) - f(\mathbf{x} \odot \mathbf{M})}{f(\mathbf{x})} &= \\ \frac{|f(\mathbf{x}) - f(\mathbf{x} \odot \mathbf{M})|}{|f(\mathbf{x})|} &= \\ \frac{||\mathbf{v} - \mathbf{v}(\mathbf{x} \odot \mathbf{M})||_p}{||\mathbf{v}||_p} & \end{aligned} \tag{9}$$

■

Remark 7 (Occlusion sensitivity analysis as a special case). The vector extension in Prop. 3 also shows that Lem. 1.1 is a special case when we wish to analyze one particular feature of $f(\mathbf{x})$, so this can be thought as a generalization of said method.

Lemma 1.2 (Representation Occlusion Sensitivity Analysis)**.** The natural extension to dealing with representations reintroduces Lem. 1.1 with the only change in the degree of responsibility calculation.

$$\tilde{\mathbf{E}}(f|\mathbf{x}) = \sum_i (1 - \mathbf{M}_i) \frac{||\mathbf{v} - \mathbf{v}(\mathbf{x} \odot \mathbf{M})||_p}{||\mathbf{v}||_p} \tag{10}$$

Proof. Analogous to Lem. 1.1. ■

Remark 8 (Representation Occlusion Sensitivity Analysis). The natural extension to dealing with representations from Lem. 1.2 reintroduces Lem. 1.1 with the only change in the degree of responsibility calculation.

1.3.1 Occlusion Sensitivity Analysis with Deep Feature Augmentation Subspaces

While Lem. 1.2 outlines a general method to determine the degree of responsibility, its sole dependence on occlusion may not capture the nuanced relationships inherent in deep learning models. Recognizing the vital role of data augmentation in training and viewing occlusion as a form of augmentation, we propose a shift from simple vector comparisons to a detailed analysis between two subspaces. A subspace here denotes a segment of the deep feature vector space defined by an occluded image and its augmentations. We strive to assess the similarity between each "occlusion subspace" and the "reference subspace", which is formed by the original image and its augmentations. Extending Lem. 1.2, we compare the size difference between two subspaces, focusing on their orthogonal degree, and measure the canonical angles between subspaces derived from varied transformations on the original and occluded images.

Proposition 4 (Subspace degree of responsibility)**.** The degree of responsibility between subspaces is the orthogonal degree between them, *i.e*.,

$$r\left(\mathbf{M}\right) = 1 - \sum_{i}^{n_c} \left(\sigma_i\left(\mathbf{V}^T\mathbf{V}_\mathbf{M}\right)\right)^2 \tag{11}$$

Proof. We extend the idea (and the notation) of difference of vectors to difference of subspaces

$$\begin{aligned} r &= \frac{||\mathbf{v} - \mathbf{v}\left(\mathbf{x} \odot \mathbf{M}\right)||_2}{||\mathbf{v}||_2} \\ &\equiv \frac{|\mathcal{V} - \mathcal{V}_\mathbf{M}|}{|\mathcal{V}|} \\ &= \frac{|\mathcal{V} - \mathcal{V}_\mathbf{M}|}{|\mathcal{V} - \mathbf{0}|} \\ &= \frac{|\mathcal{V} - \mathcal{V}_\mathbf{M}|}{1} \\ &= |\mathcal{V} - \mathcal{V}_\mathbf{M}| \\ &= 1 - simi\left(\mathcal{V}, \mathcal{V}_\mathbf{M}\right) \\ &= 1 - \sum_{i}^{n_c} \left(\sigma_i\left(\mathbf{V}^T\mathbf{V}_\mathbf{M}\right)\right)^2 \end{aligned} \tag{12}$$

where $|\mathcal{A} - \mathcal{B}|$ is a subspace distance, *i.e*., the orthogonal degree between $\mathcal{A}$ and $\mathcal{B}$ [4,5]. $\mathbf{V}$ and $\mathbf{V}_\mathbf{M} \in \mathbb{R}^{k \times d}$ are the orthonormal basis of the subspaces $\mathcal{V}$ and $\mathcal{V}_\mathbf{M}$ respectively. ■

Theorem 1.1 (Occlusion Sensitivity Analysis with Deep Feature Augmentation Subspace)**.** The natural extension to dealing with deep feature augmentation subspaces reintroduces Lem. 1.2 with the only change in the degree of responsibility calculation.

$$\tilde{\mathbf{E}}\left(f|\mathbf{x}\right) = \sum_{i}^{n_m} \left(1 - \mathbf{M}_i\right) \left(1 - \sum_{j}^{n_c} \sigma_j^2 \left(\mathbf{V}^T\mathbf{V}_{\mathbf{M}_i}\right)\right) \tag{13}$$

Proof. Analogous to Lem. 1.2. ■

2. Experiments

All models used Imagenet-1k weights provided by torchvision. Grad-CAM [11] uses Captum [8] GuidedGradCam implementation. Grad-CAM is set to track the last convolutional layer on ResNet-50 [7] and the least Batch-Normalization layer on ViT-B [3] and Swin-V2 [9]. Integrated Gradients [12] uses Captum [8] IntegratedGradients implementation. We use a null baseline, computing the integral on 128 steps with Gauss–Legendre quadrature. Occlusion Sensitivity Analysis [13,14] uses Captum [8] Occlusion, which is implemented with a sliding window of binary masks with 32 pixels and stride of 1. This is equivalent to approximately 9216 masks per image. Quantitative results for each interpreter on each model can be found at Tab. 1.

Table 1. Metric scores on ImageNet for ResNet-50, ViT-B and Swin-V2. For deletion and minimal size, lower is better (↓). For insertion, higher is better (↑). **Bold** represents the best metric for a given model, while underline is the second best.

Method	**Model**	**Minimal Size (↓)**	**Deletion (↓)**	**Insertion (↑)**
Grad-CAM [11]	ResNet-50	0.532	0.181	0.174
	ViT-B	0.512	0.340	0.415
	Swin-V2	0.502	0.374	0.279
Integrated Gradients [12]	ResNet-50	0.522	**0.086**	0.248
	ViT-B	0.541	**0.223**	0.378
	Swin-V2	0.492	**0.177**	0.393
Occlusion [14]	ResNet-50	0.255	0.278	0.456
	ViT-B	**0.343**	0.295	**0.521**
	Swin-V2	**0.155**	0.410	**0.670**
Ours	ResNet-50	**0.137**	0.291	**0.530**
	ViT-B	0.346	0.311	0.507
	Swin-V2	0.210	0.387	0.579

References

[1] H. Chockler and J. Y. Halpern. Responsibility and Blame: A Structural-Model Approach. In *Journal of Artificial Intelligence Research*, volume 22 of *Journal of Artificial Intelligence Research*, pages 93–115, Oct. 2004. 1

[2] Hana Chockler, Daniel Kroening, and Youcheng Sun. Explanations for Occluded Images. In *ICCV*, pages 1214–1223, Oct. 2021. 1, 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, June 2021. 4

[4] Kazuhiro Fukui. Subspace Methods. In *Computer Vision*, pages 1–5. Springer International Publishing, Cham, 2020. 4

[5] Kazuhiro Fukui and Atsuto Maki. Difference Subspace and Its Generalization for Subspace-Based Methods. In *PAMI*, volume 37 of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2164–2177, 2015. 4

[6] Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. In *The British Journal for the Philosophy of Science*, volume 56 of *The British Journal for the Philosophy of Science*, pages 889–911, Dec. 2005. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, June 2016. 4

[8] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. In *ICLR*, Sept. 2020. 4

[9] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *CVPR*, pages 11999–12009, June 2022. 4

[10] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *BMVC*, June 2018. 2, 3

[11] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128:336–359, Oct. 2016. 4, 5

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *ICML*, volume 70, Sydney, Australia, June 2017. jmlr.org. 4, 5

[13] Tomoki Uchiyama, Naoya Sogi, Koichiro Niinuma, and Kazuhiro Fukui. Visually explaining 3D-CNN predictions for video classification with an adaptive occlusion sensitivity analysis. In *WACV*, pages 1513–1522, Jan. 2023. 4

[14] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, volume 8689, pages 818–833, Cham, 2014. Springer International Publishing. 2, 4, 5