DEEL

DEpendable & Explainable Learning

DEEL

DEpendable & Explainable Learning

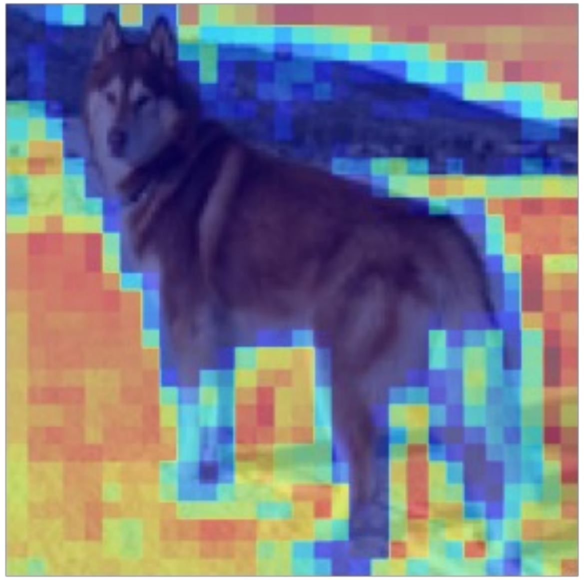
Making Sense of Dependence: Efficient Black-box Explanations Using Dependence Measure

PAUL NOVELLO, THOMAS FEL, DAVID VIGOUROUX, NEURIPS 2022

IVADO SAINT EXUPÉRY CRIAQ ANITI Institut intelligence et données

Introduction – why explainability in Deep Learning?

- Build trust in the model prediction
- Make sure the model makes a prediction for a good reason
  - Identifiy bias or spurious effects learned by a model



Introduction – why explainability in Deep Learning?

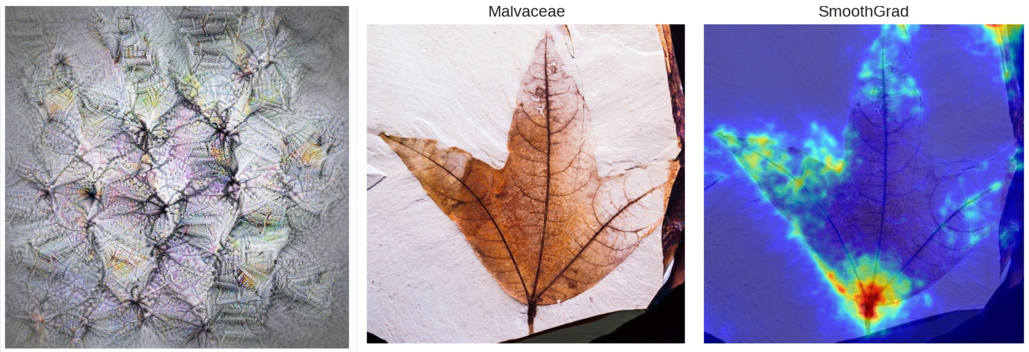
DEEL
DEpendable & Explainable Learning

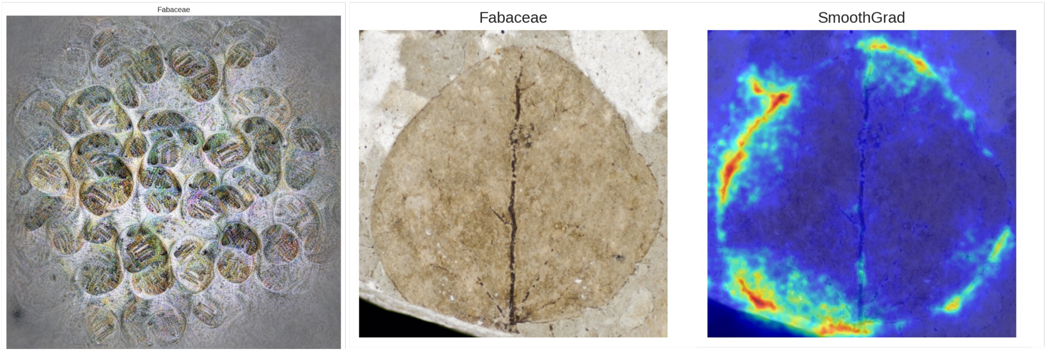
- Build trust in the model prediction
- Make sure the model makes a prediction for a good reason
  - Identifiy bias or spurious effects learned by a model
  - Understand failure cases

Introduction – why explainability in Deep Learning?

DEEL
DEpendable & Explainable Learning

- Build trust in the model prediction
- Make sure the model makes a prediction for a good reason
  - Identifiy bias or spurious effects learned by a model
  - Understand failure cases
- Pattern mining: identify patterns in data

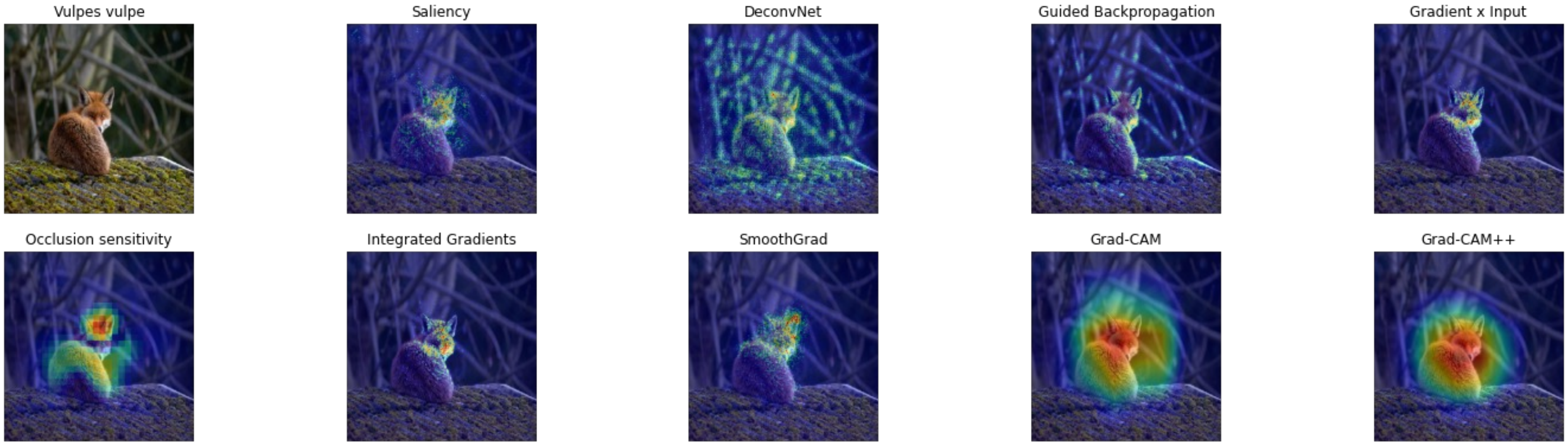




Thomas Fel – DEEL, Brown university, work in progress with Harvard university

A zoology of attribution methods

DEpendable & Explainable Learning



A zoology of attribution methods

DEependable & Explainable Learning

Saliency Maps Symonyan & al (2013)[1]

$$\Phi = \nabla f(x) \implies \phi_i = \frac{\partial f(x)}{\partial x_i}$$

In an infinitesimal neighborhood (often not feasible), what are my features that most impact the output score ?

SmoothGrad Smilkov & al (2017)[2]

$$\Phi = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I\sigma)}[\nabla f(x+\epsilon)]$$

$$\Phi = \frac{1}{N}\sum_{i=0}^{N} \nabla f(x+\epsilon)$$

As the name suggests, averages the gradient at several points corresponding to small perturbations around the point of interest.

[1] Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps
[2] SmoothGrad: removing noise by adding noise

A zoology of attribution methods

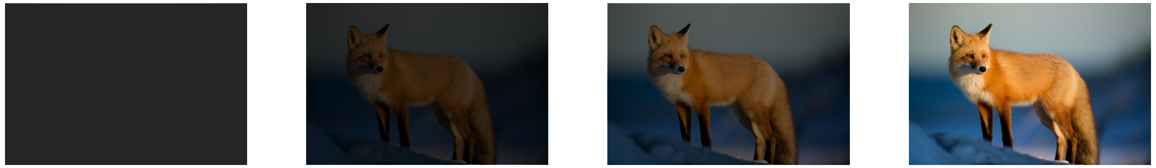
DEpendable & Explainable Learning

Integrated Gradients Sundarajan & al (2017)[1]

$$\Phi = (x - x_0)\int_0^1 \frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x} d\alpha$$

$$\Phi = (x - x_0)\frac{1}{N}\sum_{i=0}^{N} \frac{\partial f(x_0 + \frac{i}{N}(x - x_0))}{\partial x}$$

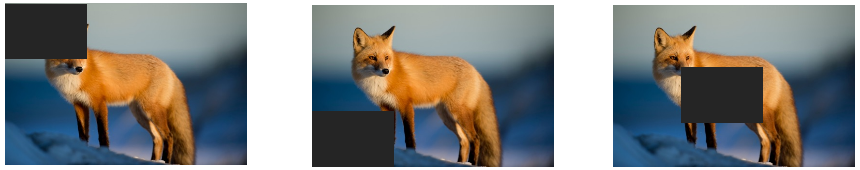
Averaging the gradient values along the path from a baseline state to the current value. The baseline state is often set to zero.



Occlusion Ancona & al (2017)[2]

$$\phi_i = f(x) - f(x_{[x_i = x_0]})$$

Sweep a patch that occludes pixels over the images, and use the variations of the model prediction to deduce critical areas.



[1] Axiomatic Attribution for Deep Networks
[2] Towards better understanding of gradient-based attribution methods for Deep Neural Networks

A zoology of attribution methods

DEel
DEpendable & Explainable Learning

RISE Petsiuk & al (2018)[1]

$$\phi_i = \mathbb{E}[f(x \odot m) | m = 1]$$

$$\phi_i = \frac{1}{\mathbb{E}[m]N} \sum_{i=0}^{N} f(x \odot m_i) \odot m_i$$

Probing the model with randomly masked versions of the input image and obtaining the corresponding outputs to deduce critical areas.

black-box state of the art

[1] RISE: Randomized Input Sampling for Explanation of Black-box Models

A zoology of attribution methods

DEpendable & Explainable Learning

And many more ...

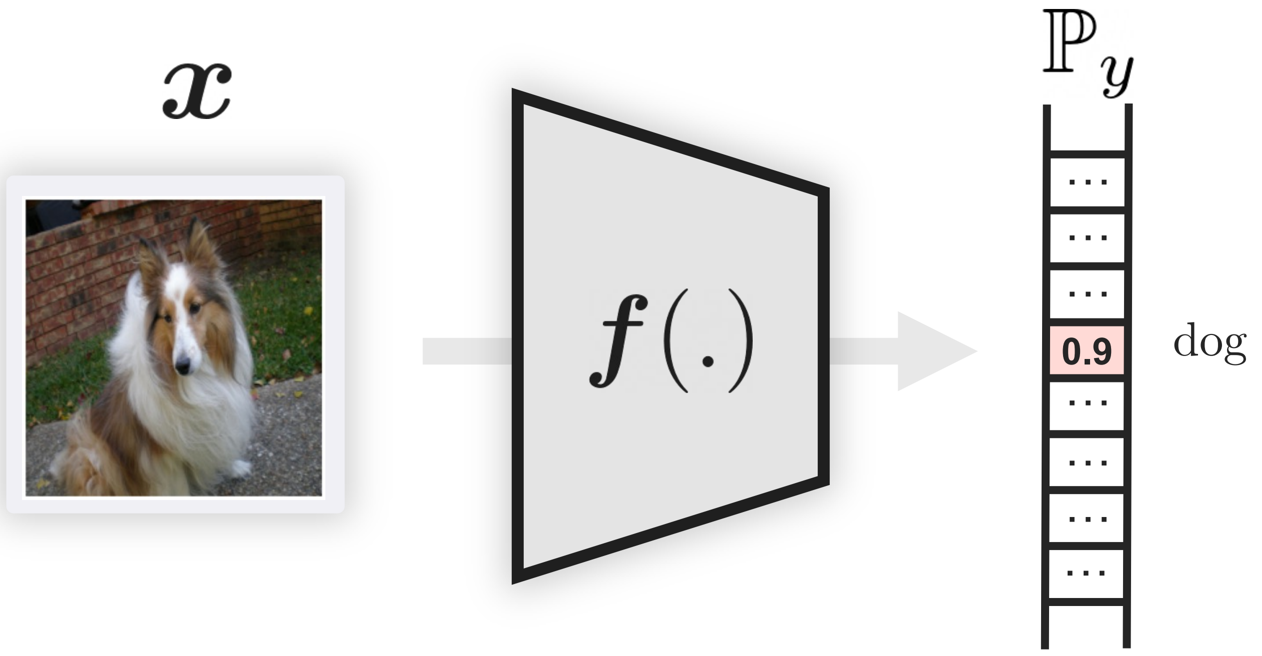
White-box

- Needs access to internal representations
- Needs a backward pass
- relatively fast

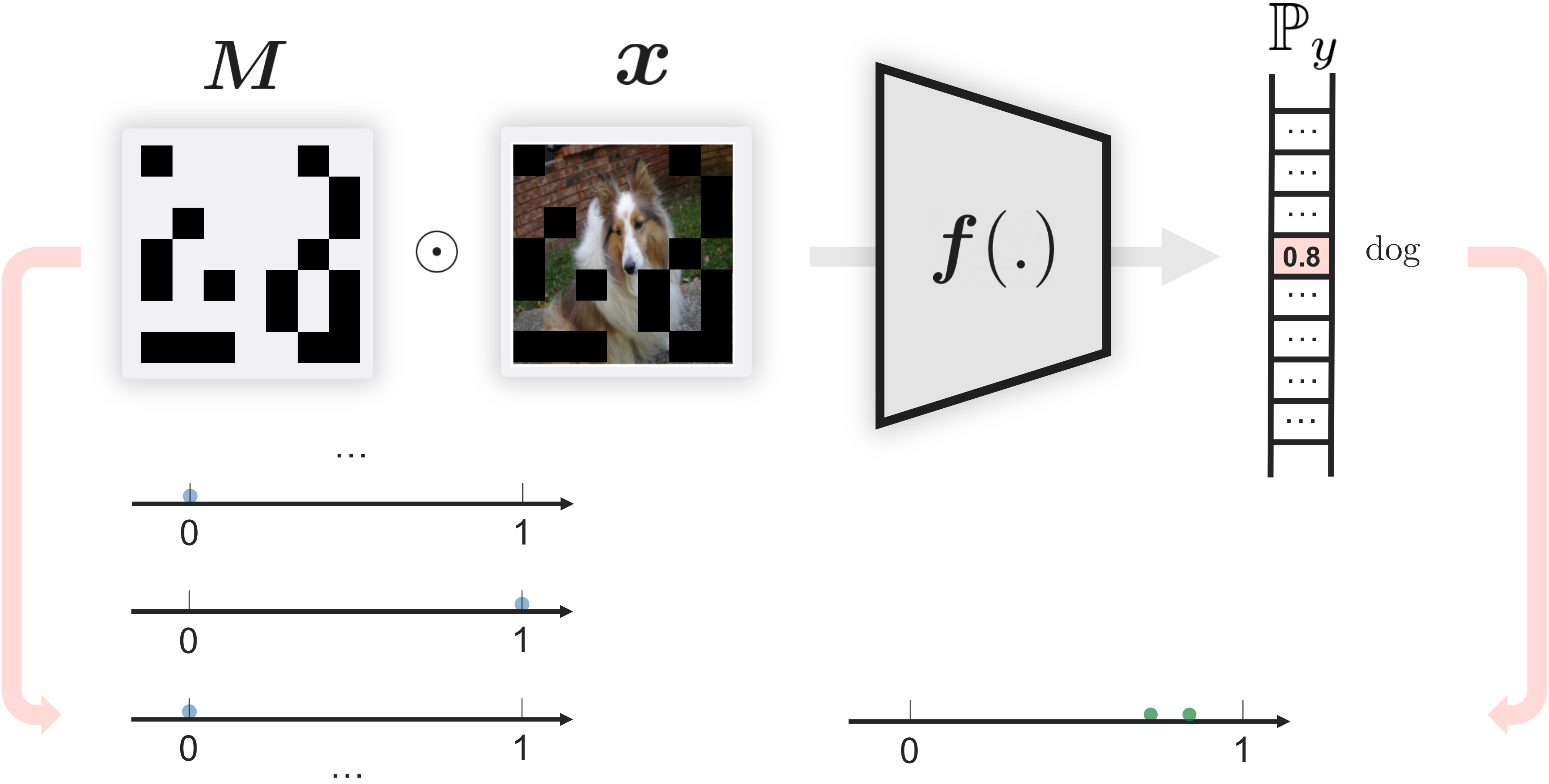
Black-box

- Only needs perturbations on the input space
- Expensive: many forward passes are required

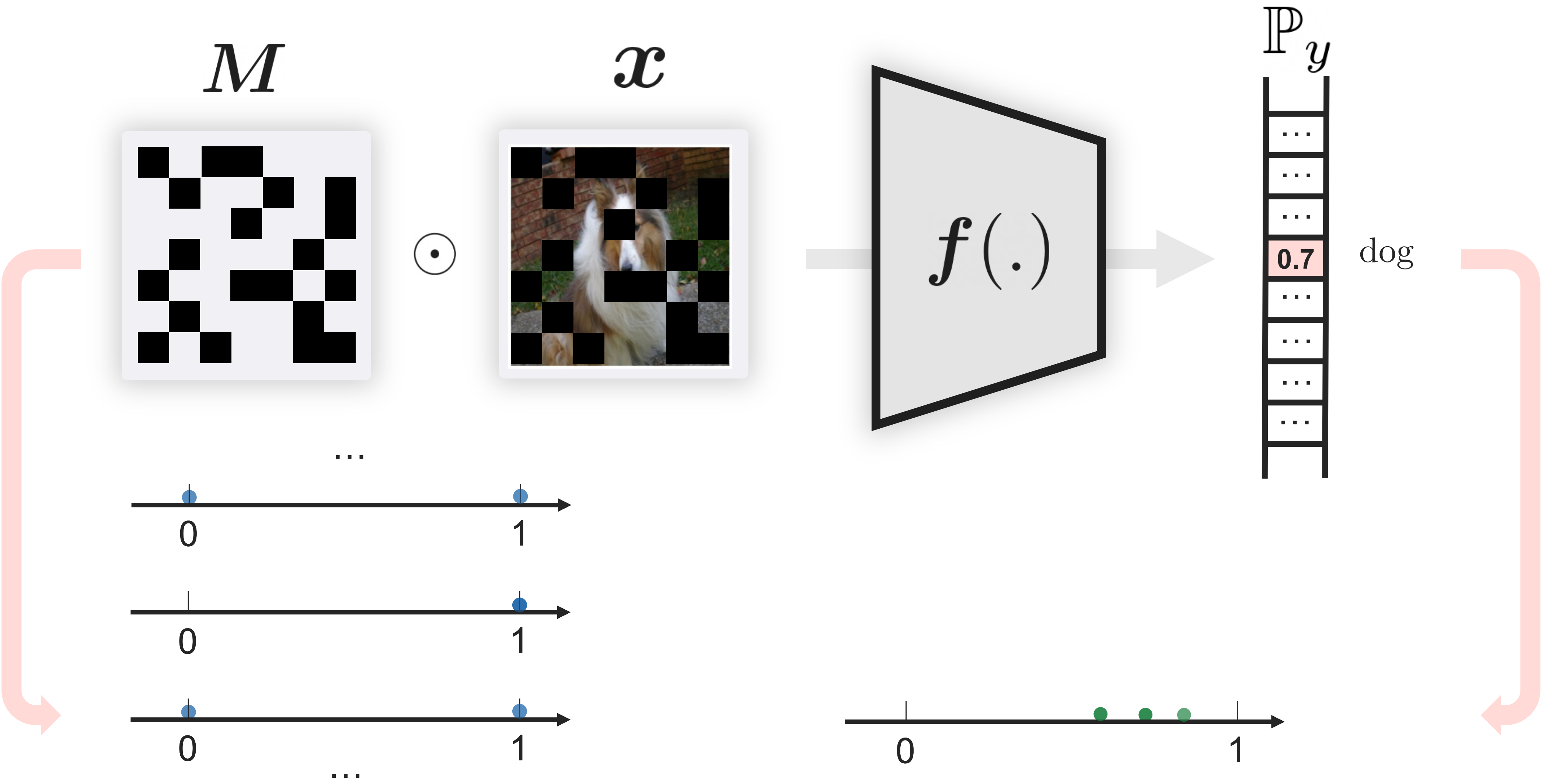
Black box: Patchwise Image Perturbation



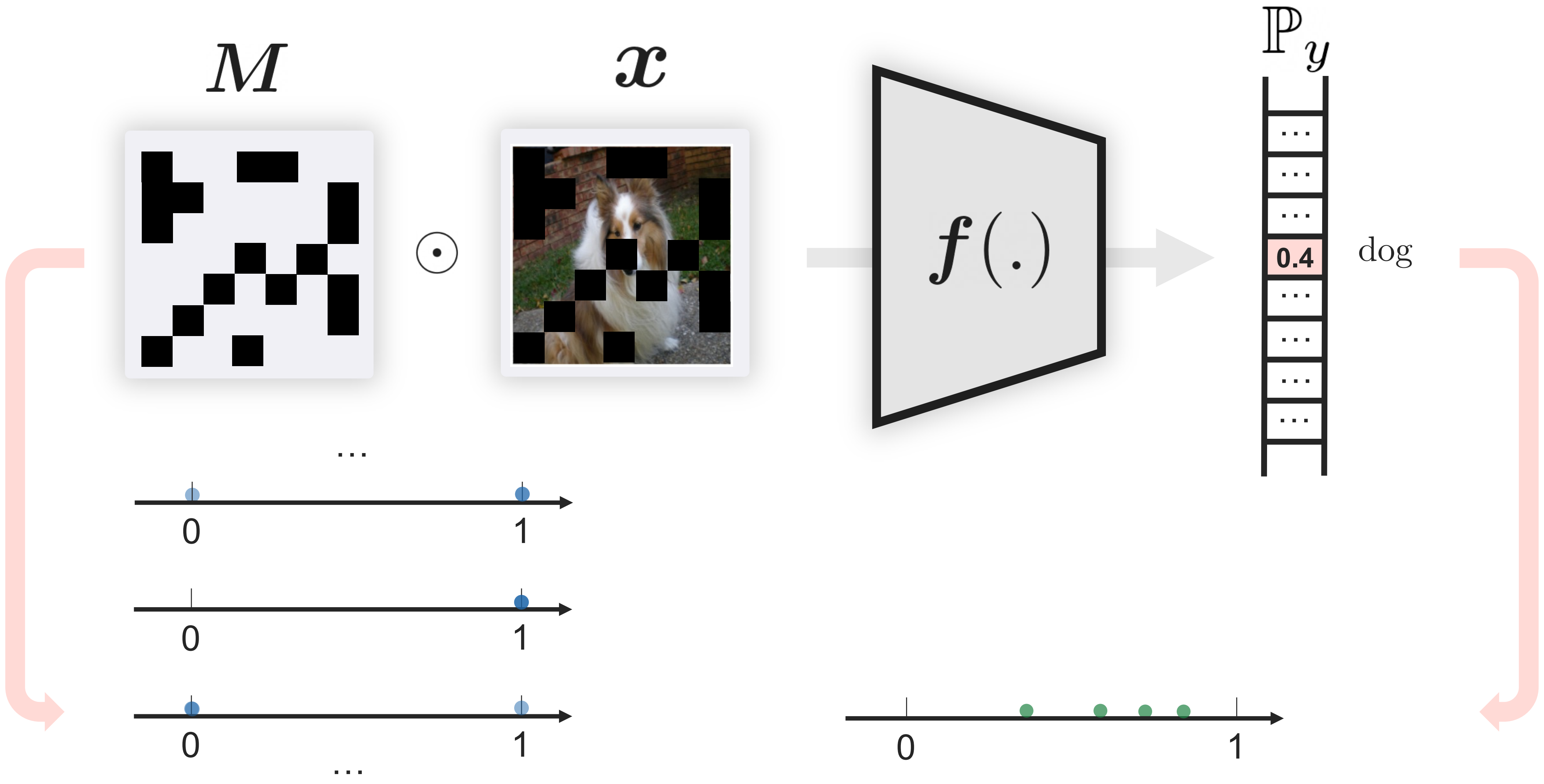
Black box: Patchwise Image Perturbation



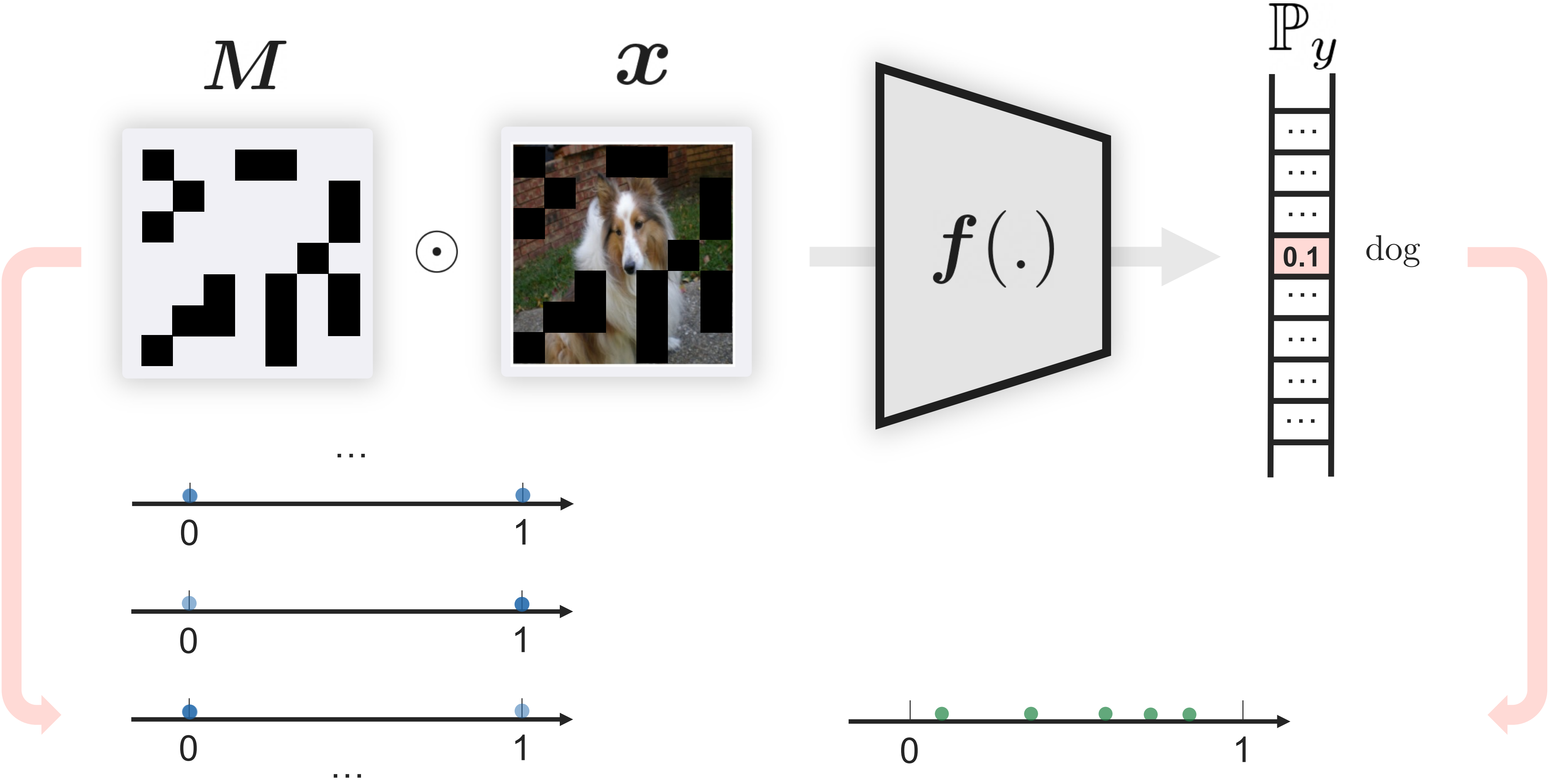
Black box: Patchwise Image Perturbation



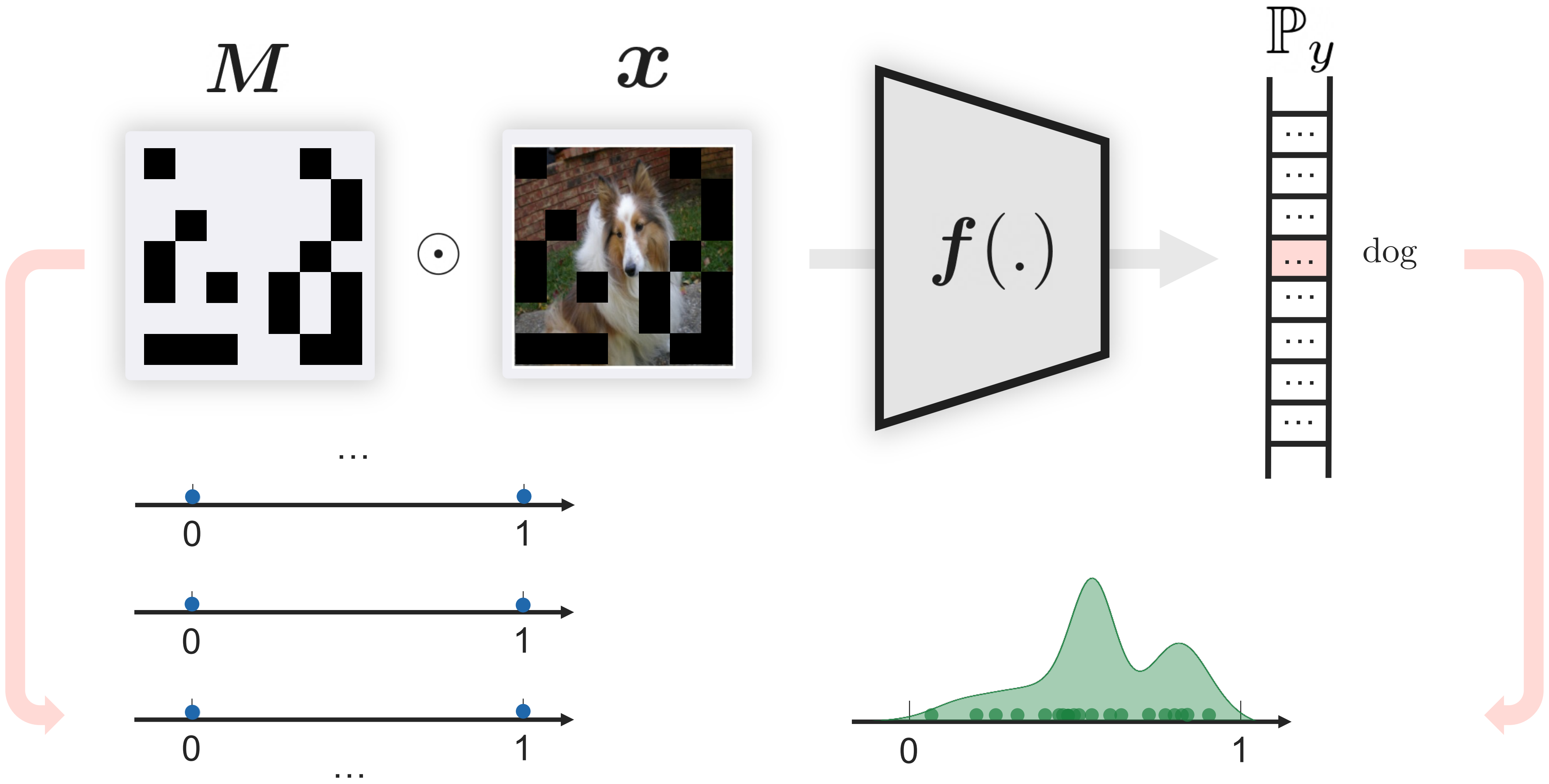
Black box: Patchwise Image Perturbation



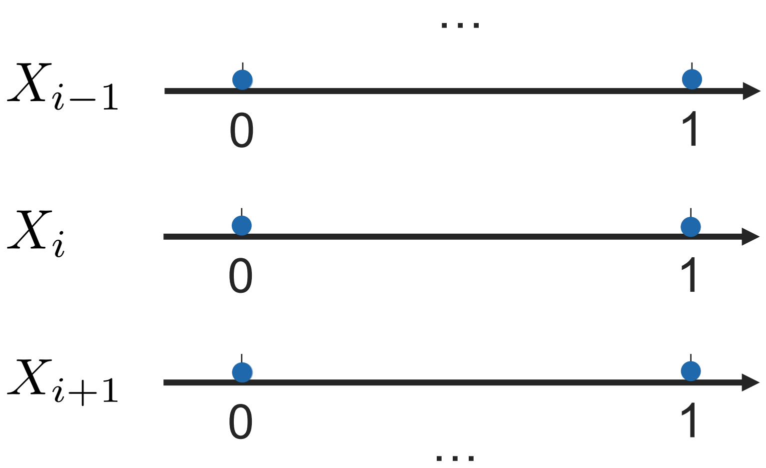
Black box: Patchwise Image Perturbation



Black box: Patchwise Image Perturbation

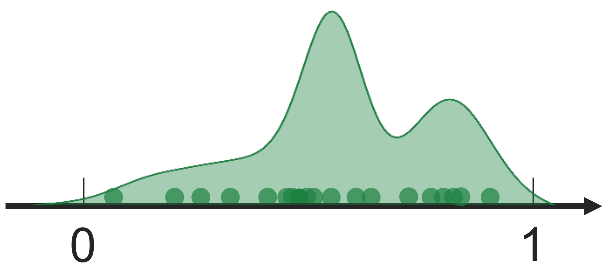


Black box: Patchwise Image Perturbation



$X_i, i \in \{1, ..., d\}$

$\mathbf{M} = \{X_1, ..., X_d\}$



Y

How can we use these p samples ?

Global sensitivity analysis

Let's consider a function $f : \begin{cases} \mathbb{R}^d & \to \mathbb{R} \\ X & \to Y = f(X) \end{cases}$

Sensitivity analysis is concerned with measuring the **sensitivity** of Y to each **input vector** $X_i, i \in \{1, ..., d\}$. **Here,** $\mathbf{M} = \{X_1, ..., X_d\}$

Global sensitivity analysis is broadly used outside A.I.

- Classical statistics
- Industrial design optimization in engineering
- Physical modeling
- ...

Global sensitivity analysis

Why Global ?

GSA (as opposed to Local SA) considers the sensitivity of Y to X_i
With respect to all its input domain.

Local

- Only considers the effect of X_i independently from one another
- Study the sensitivity of Y to small, local perturbations

Global

- Allows to draw general conclusions about the importance of a specific X_i
- Thorough analysis of the sensitivity of Y, including to interactions between X_i

Global sensitivity analysis

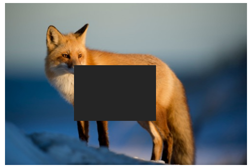
DEEL
DEpendable & Explainable Learning

Some attribution methods perform SA without knowing it !

Local

Examples:

- Occlusion



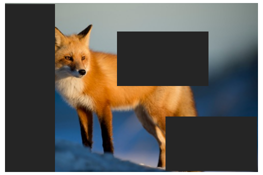
- Saliency

$$\Phi = \nabla f(x) \implies \phi_i = \frac{\partial f(x)}{\partial x_i}$$

Global

Examples:

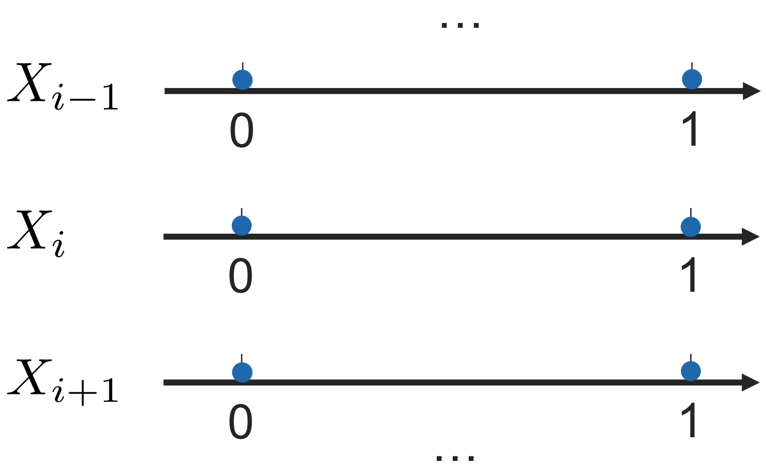
- RISE



- Sobol

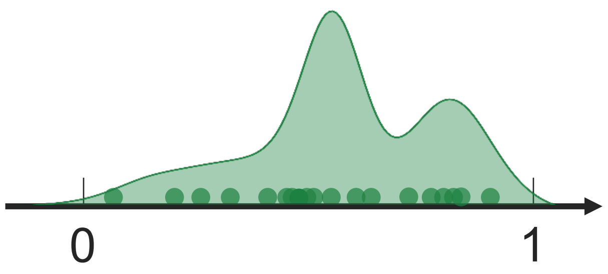
…

Flashback: Sobol attribution method



$X_i, i \in \{1, ..., d\}$

$\mathbf{M} = \{X_1, ..., X_d\}$



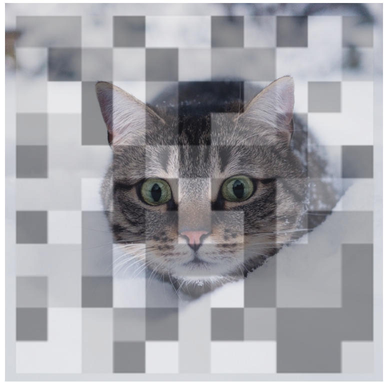
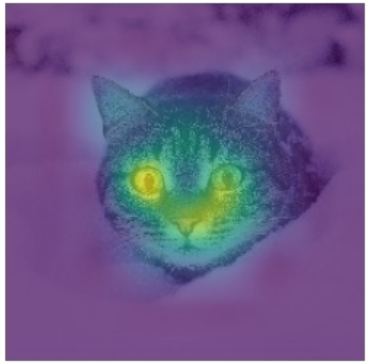
Y

Sobol method measures the importance of X_i by assessing its contribution to the variance of Y (ANOVA).

Flashback: Sobol attribution method [1]

Sobol method measures the importance of X_i by assessing its contribution to the variance of Y (ANOVA).

Each patch X_i gets an importance score which is the total Sobol index $\mathcal{S}_i$ of the corresponding X_i. Assesses the importance of X_i and of all its interactions.

$\mathbf{M} = \{X_1, ..., X_d\}$  $\longrightarrow$  $\mathcal{S}_i$

[1] Thomas Fel et al, Neurips 2021

Another approach: GSA using dependence

Idea: if Y is sensitive to X_i , then those two random variables are dependent

How to measure the dependence between two random variables ?

- Let $\mathbb{P}_{X_i}$ be the probability distribution of X_i
- Let $\mathbb{P}_Y$ be the probability distribution of Y
- Let $\|\cdot\|$ be some distance defined on probability distributions

$$\boxed{\|\mathbb{P}_{X_i}\mathbb{P}_Y - \mathbb{P}_{X_i,Y}\| = 0 \Rightarrow \mathbb{P}_{X_i}\mathbb{P}_Y = \mathbb{P}_{X_i,Y} \Rightarrow X_i \perp Y}$$

MMD and RKHS

One can measure the dependence between Y and X_i by assessing $\|\mathbb{P}_{X_i}\mathbb{P}_Y - \mathbb{P}_{X_i,Y}\|$

How to select $\|\cdot\|$ *?* *Two ingredients:*

- Reproducing Kernel Hilbert Space (**RKHS**) is a space where we can construct representations (called embeddings) of random variables.
- The Maximum Mean Discrepancy (**MMD**) is a distance defined in a Restricted Kernel Hilbert Space (RKHS). It can be used to measure the distance between the embedding of two distributions.

MMD and RKHS

- Reproducing Kernel Hilbert Space (**RKHS**) is a space where we can construct representations (called embeddings) of random variables.

  - Let $k : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a kernel
  - The embedding of $x \in \mathcal{X}$ In the RKHS $\mathcal{F}$, $\Phi : \mathcal{X} \rightarrow \mathcal{F}$ is defined by:

$$\Phi(x) := x' \rightarrow k(x, x')$$

- The Maximum Mean Discrepancy (**MMD**) is a distance defined in a Restricted Kernel Hilbert Space (RKHS). It can be used to measure the distance between the embedding of two distributions.

$$\gamma(P_{X_i}, P_Y) = MMD(P_{X_i}, P_Y) = \|\mu_{P_{X_i}} - \mu_{P_Y}\|_{\mathcal{H}}$$

Where $\mu_{P_{X_i}}$ is the mean embedding of X_i defined by $\mu_{P_{X_i}} := x' \rightarrow \int k(x, x') dP_{X_i}(x)$

Hilbert Schmidt Independence Criterion

- Let $k : \mathbb{R}^2 \to \mathbb{R}$ be a kernel used for embedding X_i, defining RKHS $\mathcal{F}$
- Let $l : \mathbb{R}^2 \to \mathbb{R}$ be a kernel used for embedding Y, defining RKHS $\mathcal{G}$
- Define a kernel $v : \mathbb{R}^4 \to \mathbb{R}^2; (x, x'), (y, y') \to k(x, x')l(y, y')$ and thus a RKHS $\mathcal{H}$

Hilbert Schmidt Independence Criterion is a measure of dependence defined on $\mathcal{H}$ by

$$\boxed{HSIC(X_i, Y) = \gamma^2(\mathbb{P}_{X_i}\mathbb{P}_Y, \mathbb{P}_{X_i,Y})}$$

Hilbert Schmidt Independence Criterion

- Let $k : \mathbb{R}^2 \to \mathbb{R}$ be a kernel used for embedding X_i
- Let $l : \mathbb{R}^2 \to \mathbb{R}$ be a kernel used for embedding Y

(defines RKHSs $\mathcal{G}$ and $\mathcal{F}$)

HSIC can be efficiently estimated using:

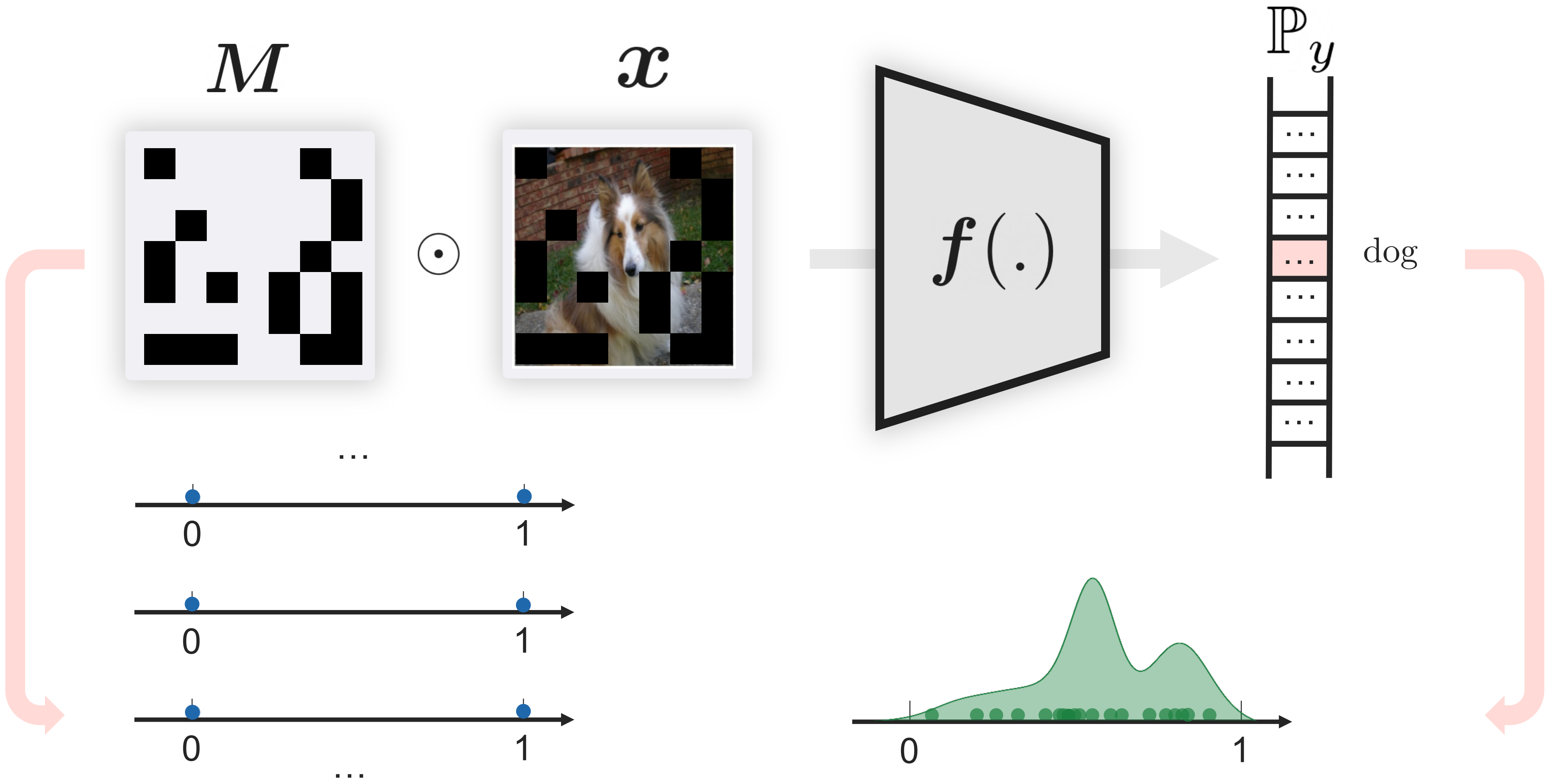
$$\boxed{\mathcal{H}^p_{X_i,Y} = \frac{1}{(p-1)^2}\,\mathrm{tr}(KHLH)}$$

where $H, L, K \in \mathbb{R}^{p\times p}$,

$$K_{jk} = k(X_i^j, X_i^k), L_{j,k} = l(Y^j, Y^k) \text{ and } H_{jk} = \delta(j=k) - p^{-1}$$

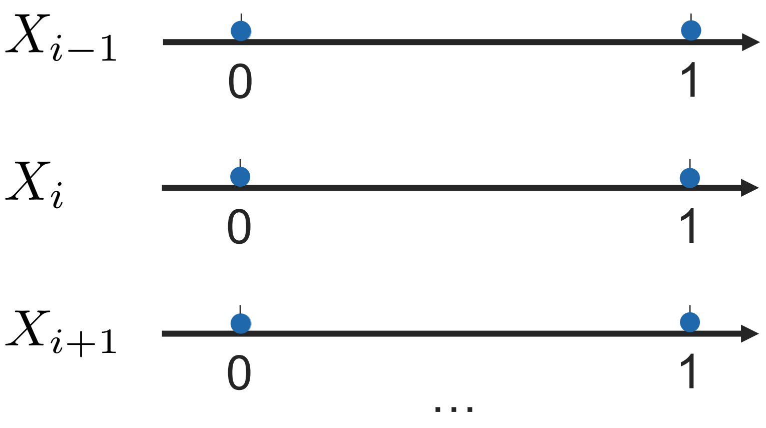
For an estimation with p samples $\{X_i^1, ..., X_i^p\}$ of X_i

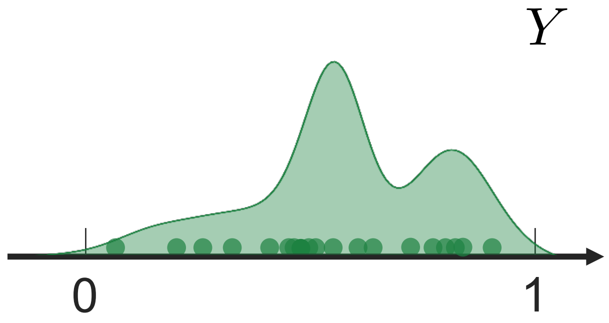
Black box: Patchwise Image Perturbation

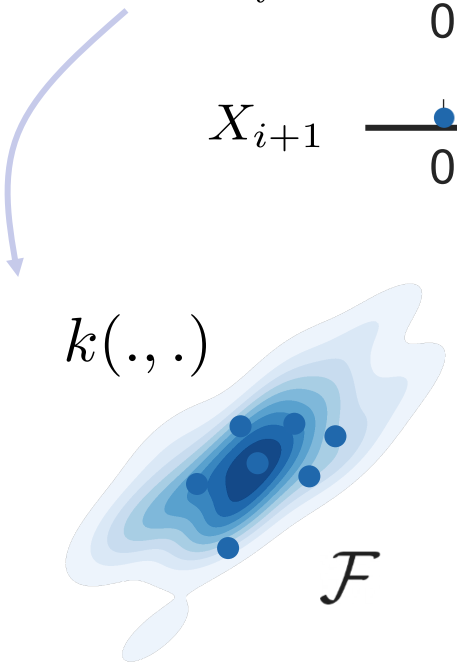


Hilbert Schmidt Independence Criterion

$$\mathbf{M} = \{X_1, ..., X_d\}$$

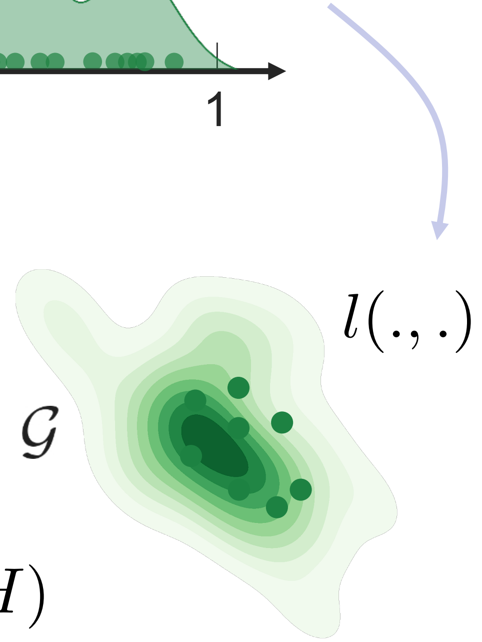






Hilbert Schmidt Independence Criterion:

$$HSIC(X_i, Y) = \gamma^2(\mathbb{P}_{X_i}\mathbb{P}_Y, \mathbb{P}_{X_i,Y})$$



Estimated with $\mathcal{H}^p_{X_i,Y} = \frac{1}{(p-1)^2}\,\mathrm{tr}(KHLH)$

Advantages of HSIC

DEpendable & Explainable Learning

Why using a different sensitivity measure ?

- The estimator $\mathcal{H}^p_{X_i,Y}$ can estimate HSIC in $\mathcal{O}(1/\sqrt{p})$ with only p samples while Sobol estimator needs $p\times(d+2)$ samples to reach the same accuracy.
- Bringing in RKHS theory opens up many research perspectives !

Practical Advantages: Efficiency

DEEL
DEpendable & Explainable Learning

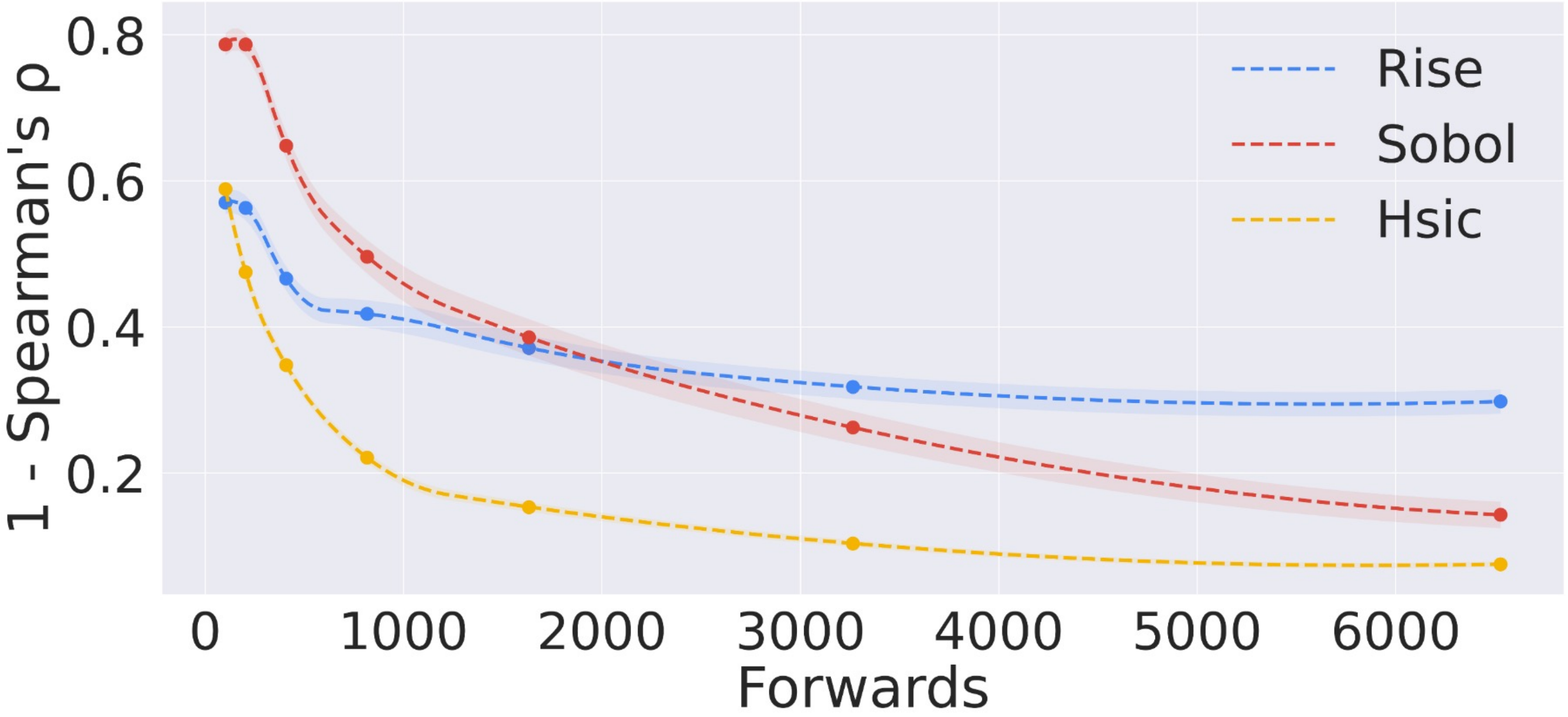
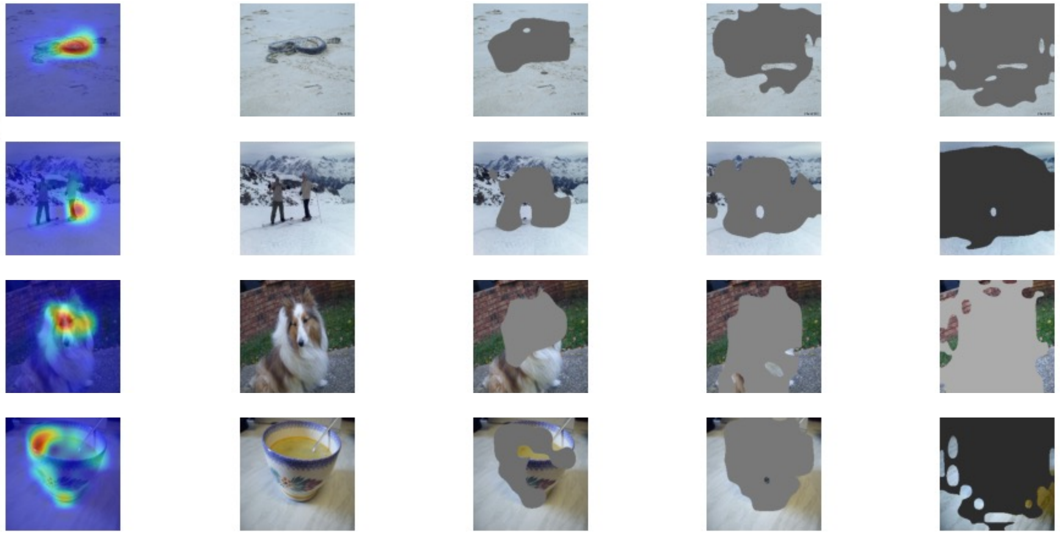


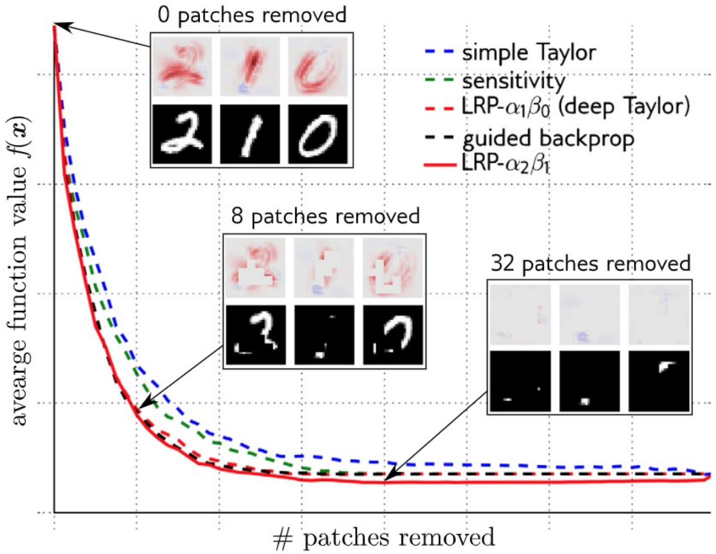
Illustration of the convergence speed of HSIC estimator against Sobol and RISE (Forwards = p)

How to evaluate the quality of explanations ?

DEEL
DEpendable & Explainable Learning

Fidelity metrics. Example: Deletion

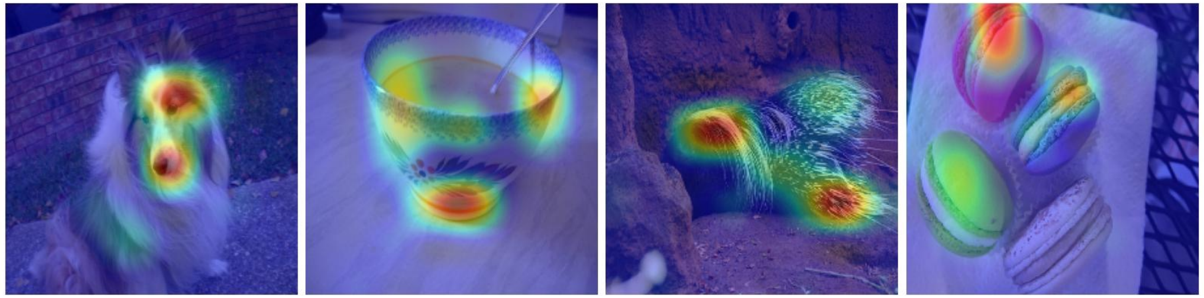




The better the explanation, the quicker the score should drop when removing important regions.

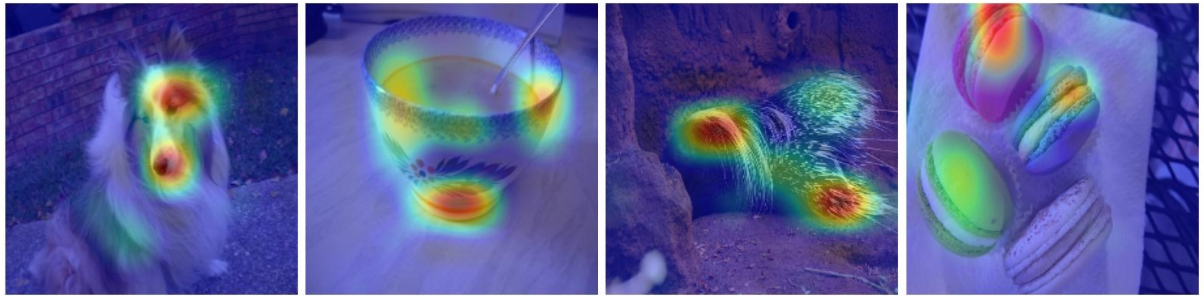
First Results: Fidelity Metrics (Deletion)

DEpendable & Explainable Learning



	Method	*ResNet50*	*VGG16*	*EfficientNet*	*MobileNetV2*
Del. (↓)					
White-box	Saliency [43]	0.158	0.120	0.091	0.113
	Grad.-Input [42]	0.153	0.116	0.084	0.110
	Integ.-Grad. [52]	0.138	**0.114**	**0.078**	0.096
	SmoothGrad [45]	0.127	0.128	0.094	**0.088**
	GradCAM++ [41]	**0.124**	0.125	0.112	0.106
	VarGrad [41]	0.134	0.229	0.224	0.097
Black-box	LIME [37]	0.186	0.258	0.186	0.148
	Kernel Shap [29]	0.185	0.165	0.164	0.149
	RISE [32]	0.114	0.106	0.113	0.115
	Sobol [11]	0.121	0.109	0.104	0.107
	$\mathcal{H}_i^p$ eff. (ours)	**0.106**	**0.100**	**0.095**	**0.094**
	$\mathcal{H}_i^p$ acc. (ours)	**0.105**	**0.099**	**0.094**	**0.093**

First Results: Fidelity Metrics (Insertion)



	Method	*ResNet50*	*VGG16*	*EfficientNet*	*MobileNetV2*
Ins. (↑)					
White-box	Saliency [43]	0.357	0.286	0.224	0.246
	Grad.-Input [42]	0.363	0.272	0.220	0.231
	Integ.-Grad. [52]	0.386	0.276	0.248	0.258
	SmoothGrad [45]	0.379	0.229	0.172	0.246
	GradCAM++ [41]	0.497	**0.413**	**0.316**	0.387
	VarGrad [41]	**0.527**	0.241	0.222	**0.399**
Black-box	LIME [37]	0.472	0.273	0.223	0.384
	Kernel Shap [29]	0.480	0.393	0.367	0.383
	RISE [32]	**0.554**	**0.485**	**0.439**	**0.443**
	Sobol [11]	0.370	0.313	0.309	0.331
	$\mathcal{H}_i^p$ eff. (ours)	0.470	0.387	0.357	0.381
	$\mathcal{H}_i^p$ acc. (ours)	0.481	0.395	0.366	0.392

DEpendable & Explainable Learning

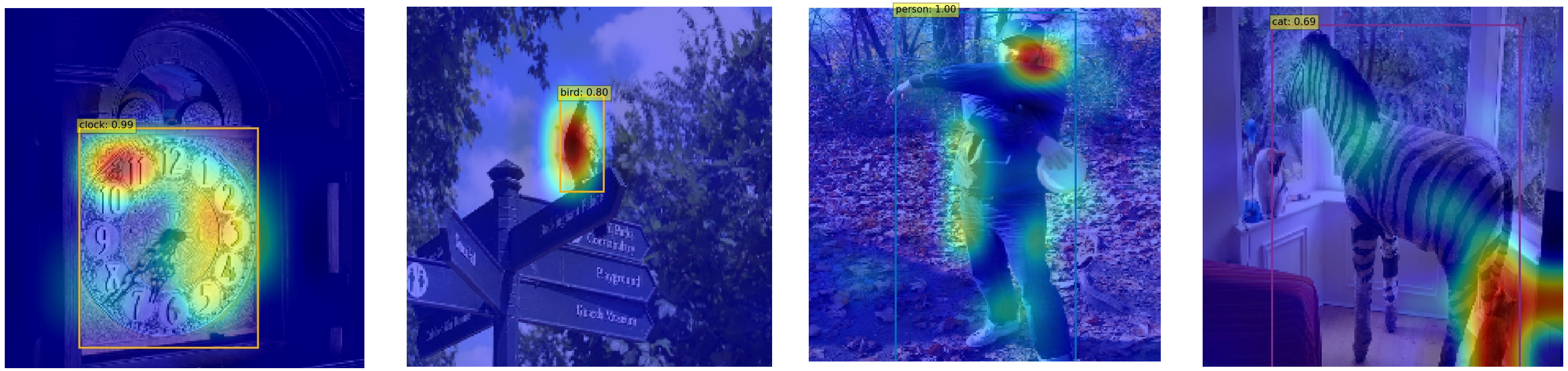
Practical Advantages: Efficiency

DEEL
DEpendable & Explainable Learning

	Method	*ResNet50*	*VGG16*	*EfficientNet*	*MobileNetV2*	Exec. time (s)
Del. (↓)						
White-box	Saliency [43]	0.158	0.120	0.091	0.113	0.360
	Grad.-Input [42]	0.153	0.116	0.084	0.110	0.023
	Integ.-Grad. [52]	0.138	**0.114**	**0.078**	0.096	**1.024**
	SmoothGrad [45]	0.127	0.128	0.094	**0.088**	0.063
	GradCAM++ [41]	**0.124**	0.125	0.112	0.106	0.127
	VarGrad [41]	0.134	0.229	0.224	0.097	0.097
Black-box	LIME [37]	0.186	0.258	0.186	0.148	6.480
	Kernel Shap [29]	0.185	0.165	0.164	0.149	4.097
	RISE [32]	0.114	0.106	0.113	0.115	8.427
	Sobol [11]	0.121	0.109	0.104	0.107	5.254
	$\mathcal{H}_i^p$ eff. (ours)	**0.106**	**0.100**	**0.095**	**0.094**	**0.956**
	$\mathcal{H}_i^p$ acc. (ours)	**0.105**	**0.099**	**0.094**	**0.093**	1.668

Explanations of Bounding Boxes

DEEL
DEpendable & Explainable Learning



Method	Deletion (↓)	Insertion (↑)	μFidelity (↑)	Exec. time (s)
D-RISE [36]	0.074	0.634	0.442	155
Kernel Shap. [32]	**0.070**	0.646	0.476	192
$\mathcal{H}_i^p$ (ours)	0.088	**0.658**	**0.568**	**34**

Explanation of Yolov4 on COCO dataset

Shortcoming of HSIC: interactions

Let $A = \{l_1, ..., l_{|A|}\} \in \mathcal{P}_d$ i.e. a subset of $\{1, ..., d\}$

For Sobol indices, we have

$$\mathcal{S}_A = \sum_{B \subset A} (-1)^{|A|-|B|} \frac{Var\mathbb{E}(Y|X_B)}{VarY}$$

When $A = \{i, j\}$, $\mathcal{S}_A = \mathcal{S}_{i,j}$ can be simply obtained with

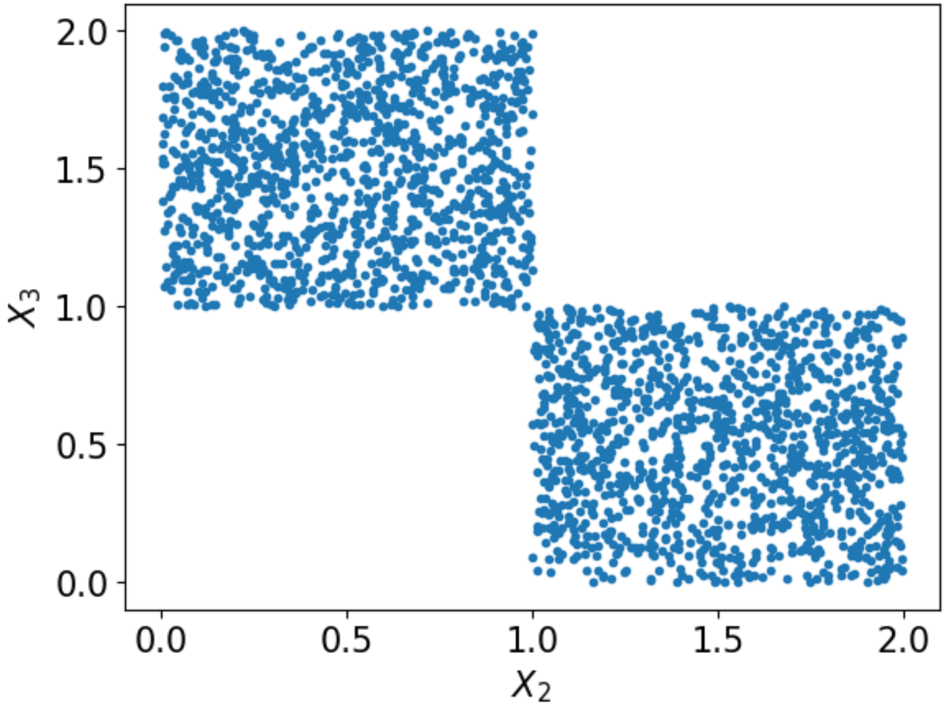
$$\mathcal{S}_{i,j} = \boxed{\text{Not possible with HSIC}} - \mathcal{S}_i - \mathcal{S}_j$$

(but expensive...)

Why considering interactions?

DEEL
DEpendable & Explainable Learning

$$Y = f(X_1, X_2, X_3) = \begin{cases} 1 & \text{if } X_1 \in [0,1], X_2 \in [1,2], X_3 \in [0,1], \\ 1 & \text{if } X_1 \in [0,1], X_2 \in [0,1], X_3 \in [1,2], \\ 0 & \text{otherwise.} \end{cases}$$



Why considering interactions?

DEpendable & Explainable Learning

$$Y = f(X_1, X_2, X_3) = \begin{cases} 1 & \text{if } X_1 \in [0,1], X_2 \in [1,2], X_3 \in [0,1], \\ 1 & \text{if } X_1 \in [0,1], X_2 \in [0,1], X_3 \in [1,2], \\ 0 & \text{otherwise.} \end{cases}$$

- X_1 is clearly important to explain Y
- X_2 and X_3 are more difficult to interpret:

$$HSIC(\mathrm{x}_2, \mathrm{y}) = 0 \text{ and } HSIC(\mathrm{x}_3, \mathrm{y}) = 0$$

...whereas they clearly have an effect on. Y !

We have to look at interactions

ANOVA-like orthogonal decomposition of HSIC

DEEL
DEpendable & Explainable Learning

In [1], an ANOVA like decomposition property is constructed for HSIC:

Let $A = \{l_1, ..., l_{|A|}\} \in \mathcal{P}_d$ i.e. a subset of $\{1, ..., d\}$

$$HSIC_A = \sum_{B \subset A} (-1)^{|A|-|B|} HSIC(X_B, Y)$$

When $A = \{i, j\}$, $HSIC_A = HSIC_{i,j}$ can be simply obtained with

$$\boxed{HSIC_{i,j} = HSIC\big((X_i, X_j), Y\big) - HSIC(X_i, Y) - HSIC(X_j, Y)}$$

...for a certain choice of kernel k_A

[1] Da Veiga, 2021

ANOVA-like orthogonal decomposition of HSIC

...for a certain choice of kernel k_A

$$k_A(X_A, X'_A) = \prod_{i \in A}(1 + k_0(X_i, X'_i))$$

with $$k_0(X, X') = k(X, X') - \boxed{\frac{\int k(X,t)dP(t) \int k(X',t)dP(t)}{\int \int k(s,t)dP(s)dP(t)}}$$

Difficult to compute

Proposition: if the kernel is constructed as

$$k_A(X_A, X'_A) = \prod_{i \in A}(1 + k_0(X_i, X'_i))$$

with $$k_0(X, X') = k(X, X') - \frac{\int k(X,t)dP(t) \int k(X',t)dP(t)}{\int \int k(s,t)dP(s)dP(t)}$$

$$\boldsymbol{k_0(X, X') = \delta(X = X') - \frac{1}{2}}$$

Interactions can be computed using orthogonal decomposition !

ANOVA-like orthogonal decomposition of HSIC

Proposition: if the kernel is constructed as

$$k_A(X_A, X'_A) = \prod_{i \in A}(1 + k_0(X_i, X'_i))$$

with $$k_0(X, X') = \delta(X = X') - \frac{1}{2}$$

Interactions can be computed using orthogonal decomposition !

Advantages of HSIC

DEpendable & Explainable Learning

Why using a different sensitivity measure ?

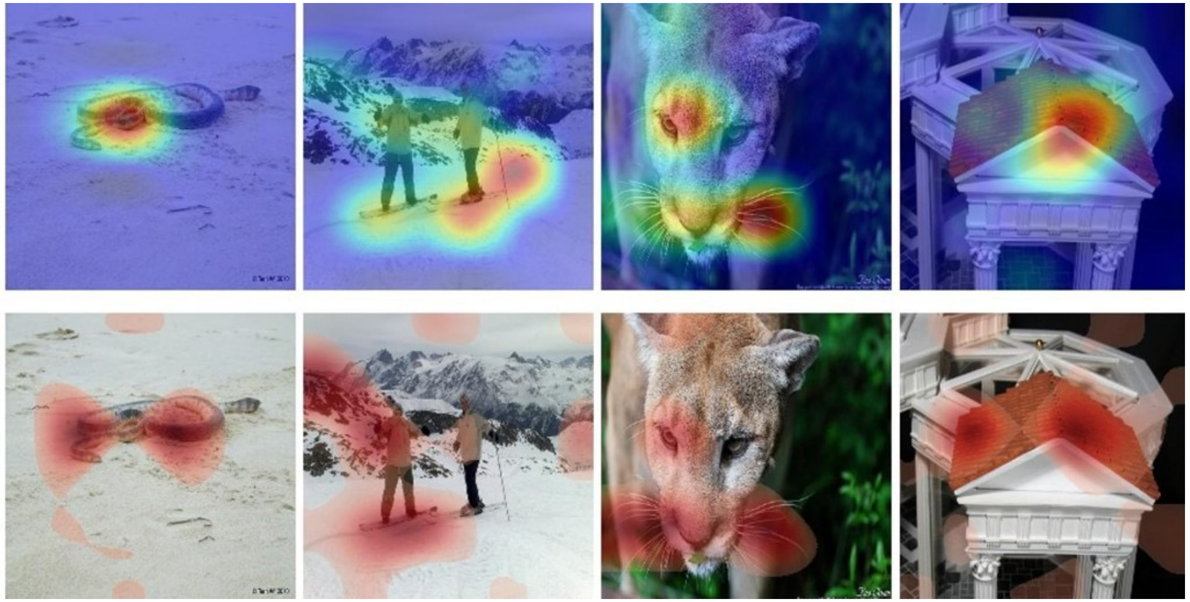
- The estimator $\mathcal{H}^p_{X_i,Y}$ can estimate HSIC in $\mathcal{O}(1/\sqrt{p})$ with only p samples while Sobol estimator needs $p\times(d+2)$ samples to reach the same accuracy.
- Bringing in RKHS theory opens up many research perspectives !

Example: now, can assess pairwise interactions !

Practical Advantages: ANOVA decomposition

$\mathcal{H}_i^p$

$\mathcal{H}_{i\times j}^p$



Conclusion and take away

DEEL
DEpendable & Explainable Learning

Context

- Black box attribution methods based on patch perturbations are versatile and convenient ways of obtaining explanations
- They suffer from high computational costs because they need many forward passes
- Global sensitivity analysis is a promising approach to exploit these perturbation
- The current SOTA GSA based attribution method uses analysis of variance with Sobol indices.

We propose to use GSA based on dependence measures (HSIC)

- Needs less forward to obtain good explanations
- Theoretical advantages of RKHS
- Can assess patch-wise interactions