One Wave to Explain them All: A Unifying Perspective on Feature Attribution

Gabriel Kasmi, Amandine Brunetto, Thomas Fel, Jayneel Parekh

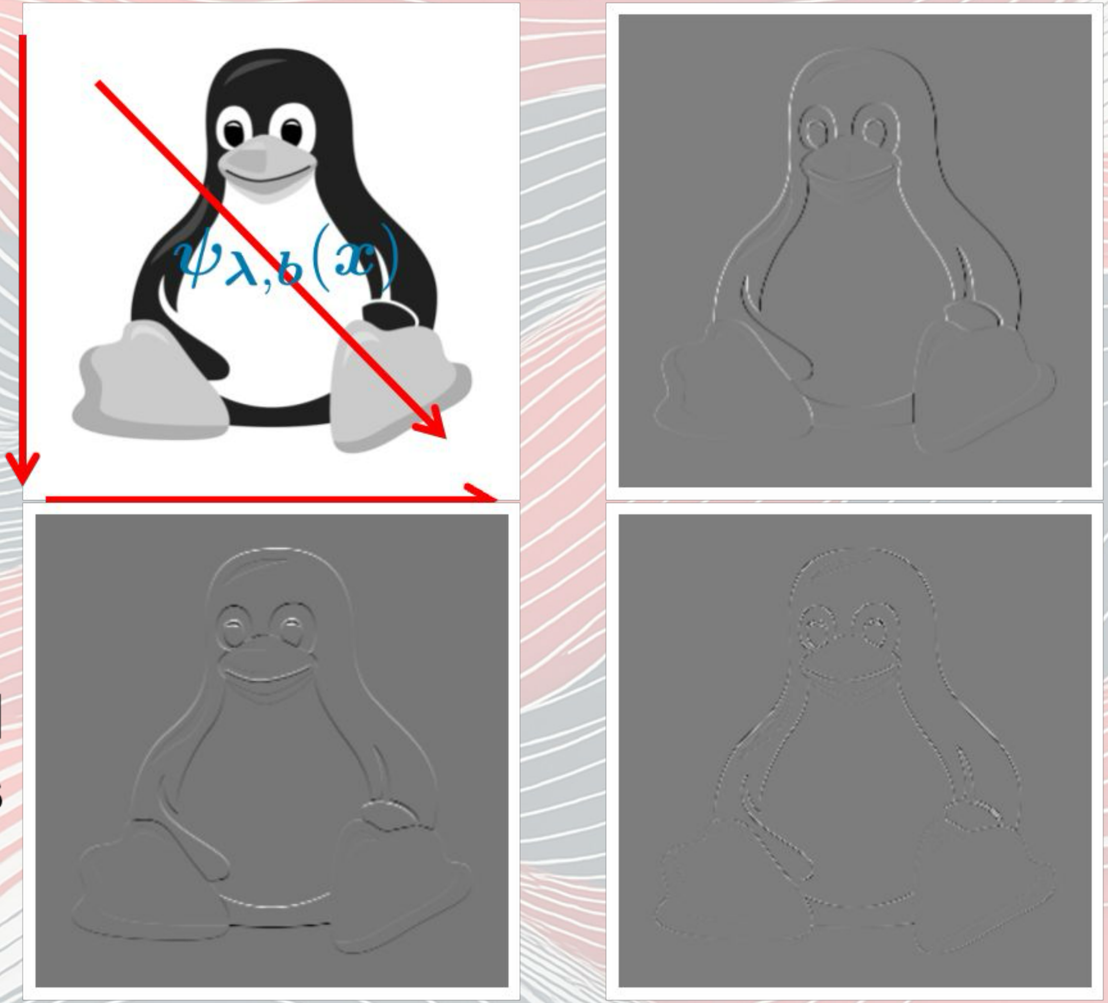
Explainable AI **aims to improve the transparency** of deep learning models.

Feature attribution: **quantify the importance of a given input feature** in the model's prediction.

For **high-dimensional data** (images, sounds, volumes) **pixel-based heatmaps**.

Pixels: **intuitive** for images but **not well-suited** for other modalities.

Pixels (or superpixels) provide **only spatial information**, but do not capture information such as frequency content.



Horizontal coefficients

Vertical coefficients

Diagonal coefficients

1-level dyadic transform

Approximation coefficients

Horizontal coefficients

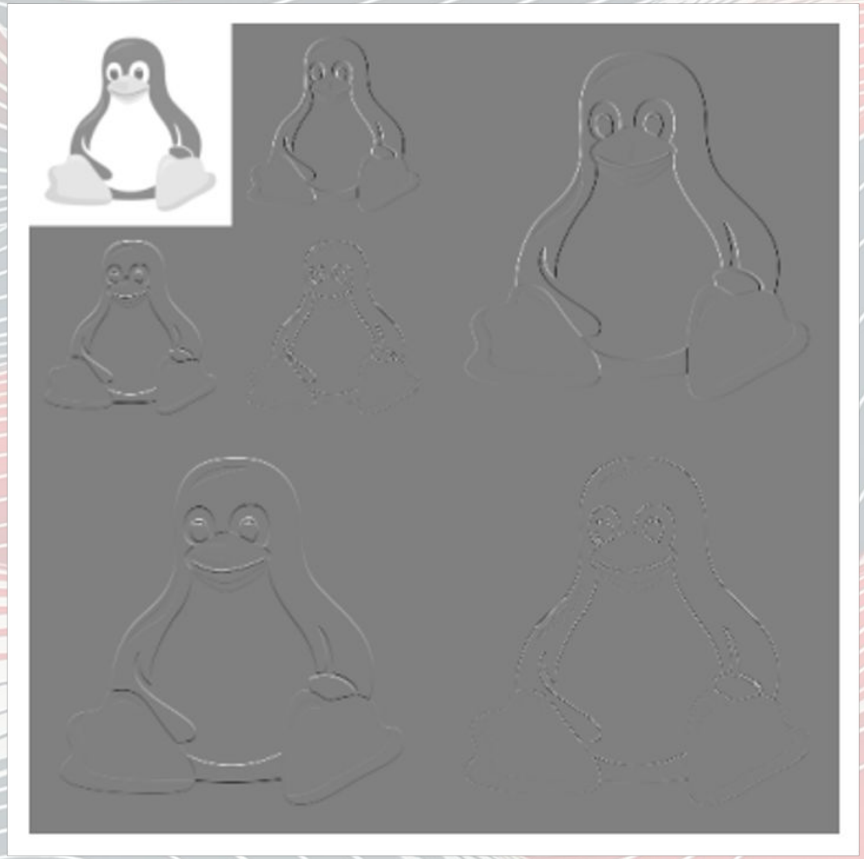
Vertical coefficients

Diagonal coefficients

2-level dyadic transform

Approximation coefficients

Horizontal coefficients



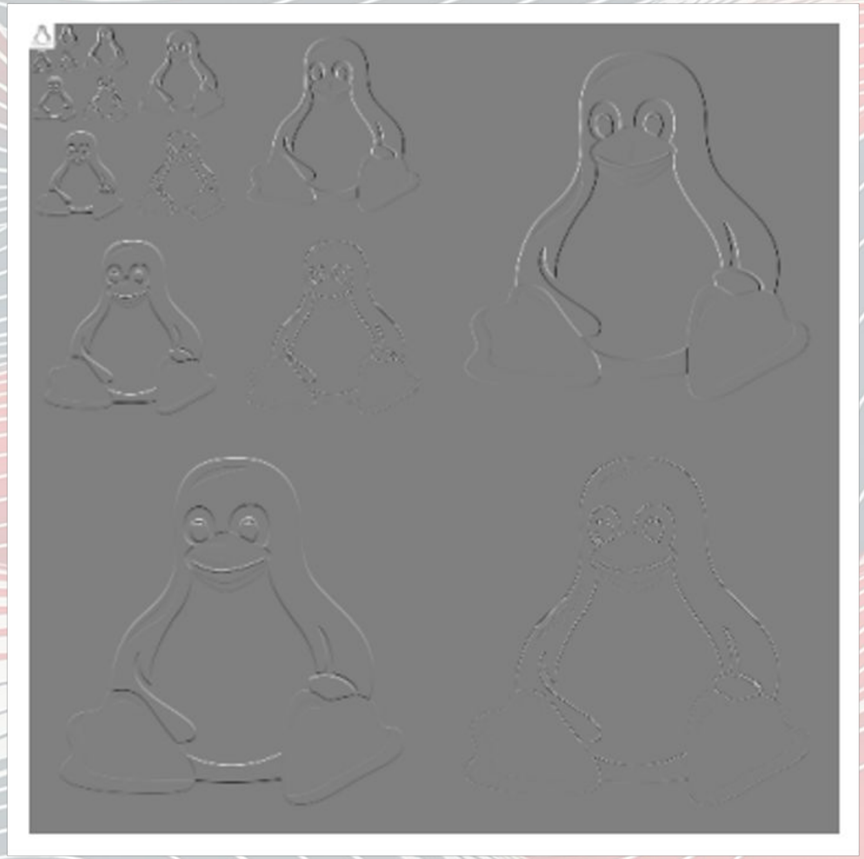
Vertical coefficients

Diagonal coefficients

n-level dyadic decomposition

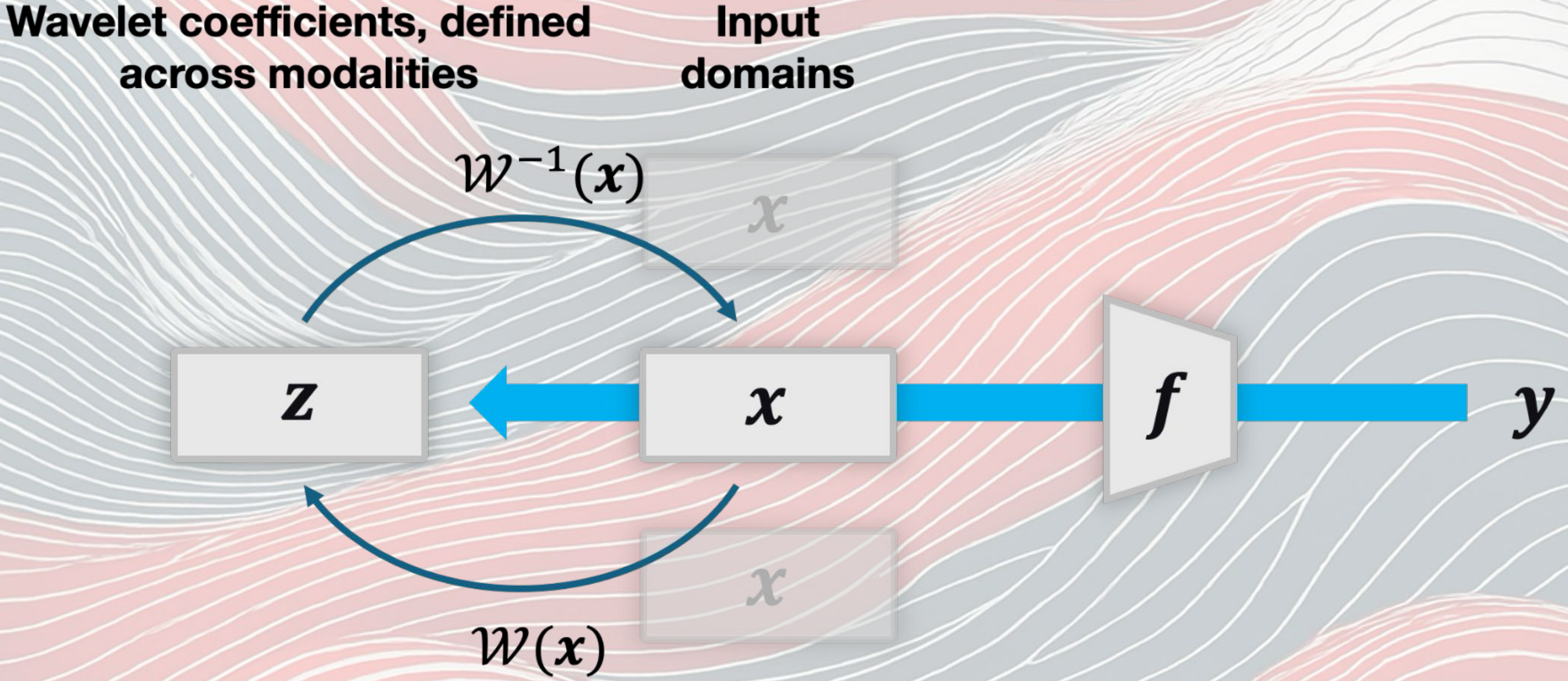
Approximation coefficients

Horizontal coefficients



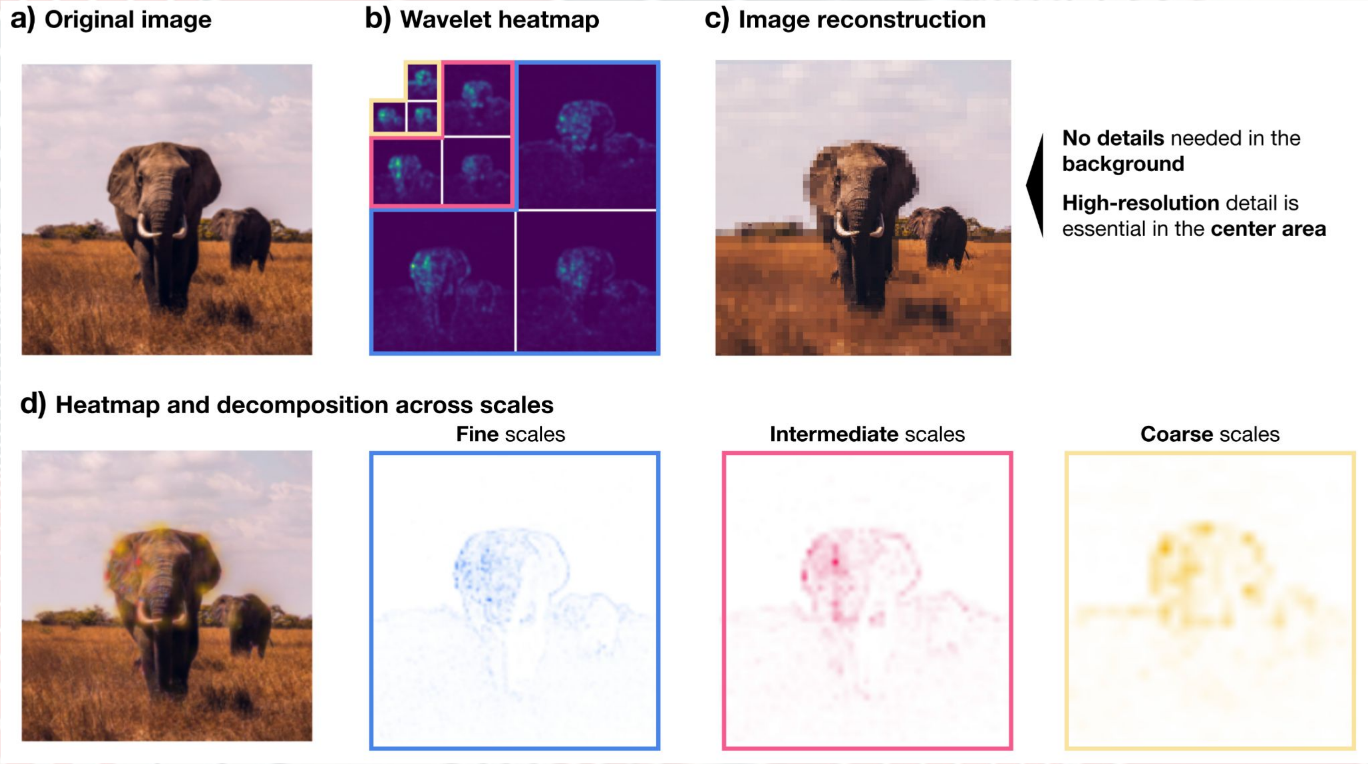
Vertical coefficients

Diagonal coefficients



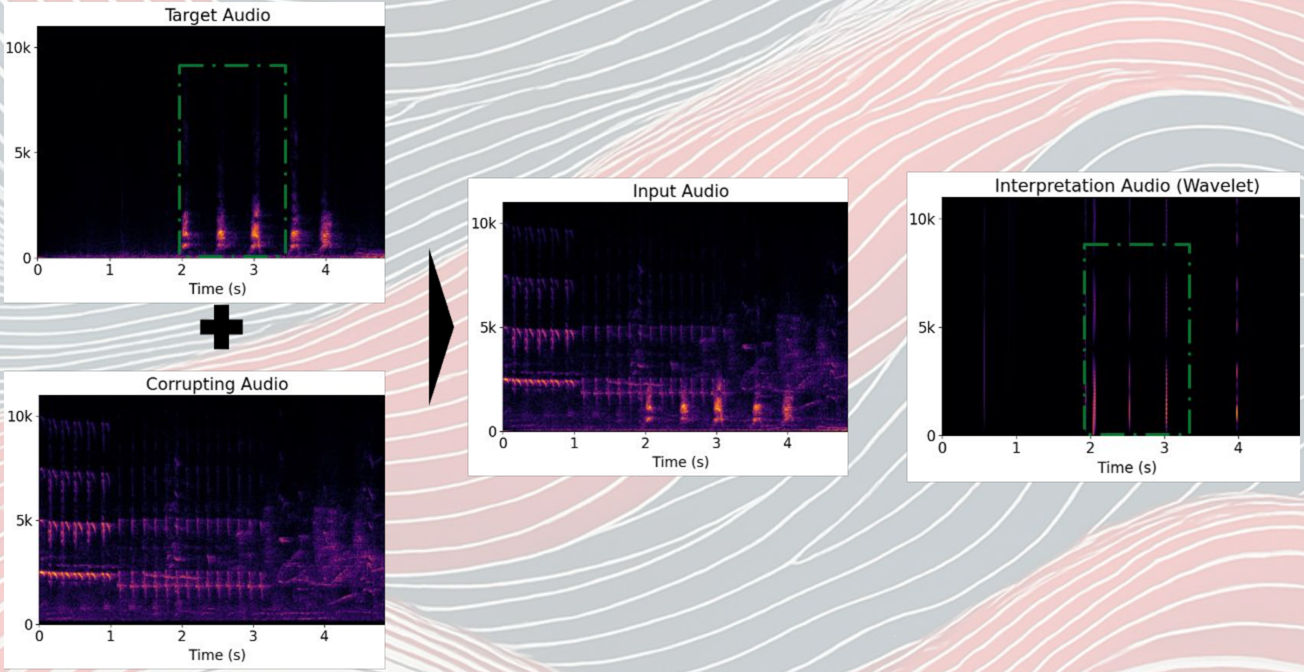
$\left|\frac{\partial \boldsymbol{f}_c(\boldsymbol{x})}{\partial z}\right|$ Computation of the gradients with respect to the **wavelet coefficients** of the input modality

More informative feature attribution



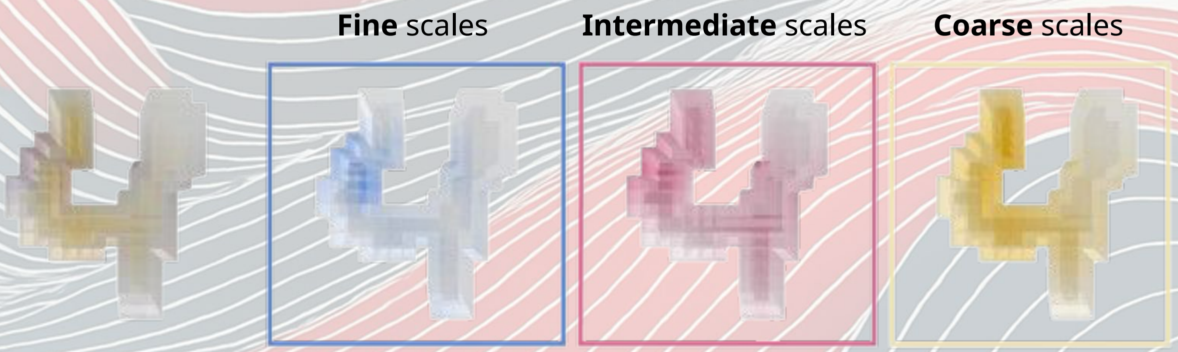
MINES PARIS | PSL
Rte
Kempner INSTITUTE
HARVARD UNIVERSITY
ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE
ICML International Conference On Machine Learning

Overlap experiment: WAM eliminates the corrupting audio from the interpretation

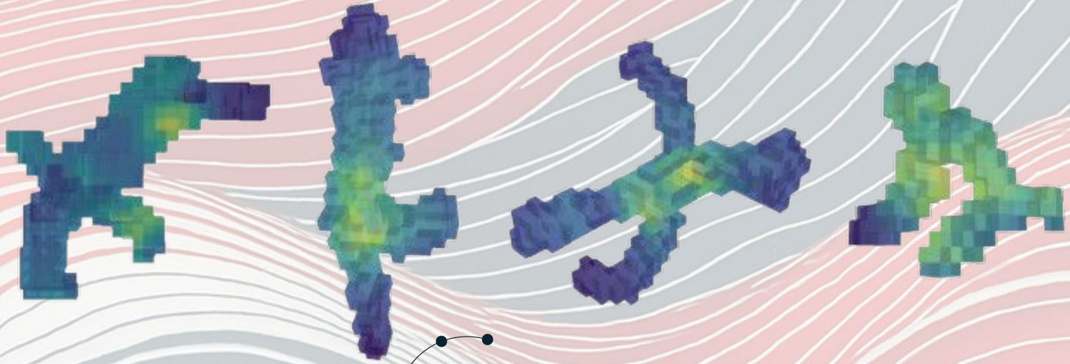


Corruption Audio

Decomposition of different scales on 3D MNIST examples



Heatmaps on VesselMNIST **volumes**



MINES PARIS | PSL
Rte
Kempner INSTITUTE
HARVARD UNIVERSITY
ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE
ICML International Conference On Machine Learning

Quantitative evaluation

	Audio			Images			Volumes		
Model	*ResNet*			*EfficientNet*			*3D Former*		
Dataset	*ESC-50*			*ImageNet*			*AdrenalMNIST3D*		
	Ins (↑)	Del (↓)	Faith (↑)	Ins(↑)	Del (↓)	Faith (↑)	Ins (↑)	Del (↓)	Faith (↑)
Integrated Gradients	0.267	**0.047**	**0.264**	0.113	0.113	0.000	0.666	0.743	-0.077
SmoothGrad	0.251	0.067	0.184	0.129	0.119	0.010	0.680	0.731	-0.051
GradCAM	0.274	0.201	0.072	0.364	0.303	0.061	0.689	0.744	-0.055
Saliency	0.220	0.154	0.066	0.148	0.140	0.008	0.751	0.742	0.009
WAM_{IG} (ours)	0.436	0.260	0.176	**0.447**	**0.049**	**0.370**	**0.719**	**0.621**	**0.098**
WAM_{SG} (ours)	**0.449**	0.252	0.197	0.419	0.097	0.350	0.718	0.648	0.070

WAM **outperforms existing methods** across a **wide range** of **metrics**, **model topologies** and **datasets** in the **audio**, **images** and **volume** cases.

Conclusions and perspectives

We **expand** gradient-based **feature attribution** to the **wavelet domain**, a **unified** and **more expressive domain**.

Future works could **expand our approach** to non smooth or non regular modalities such as **text** or **point cloud data**.

More broadly, our work discusses the **choice of the domain** over which **features** are defined.

Meet us at poster session 2 !

ICML
International Conference
On Machine Learning




MINES PARIS | PSL

Rte

Kempner INSTITUTE | HARVARD UNIVERSITY

ISIR
INSTITUT
DES SYSTÈMES
INTELLIGENTS
ET DE ROBOTIQUE