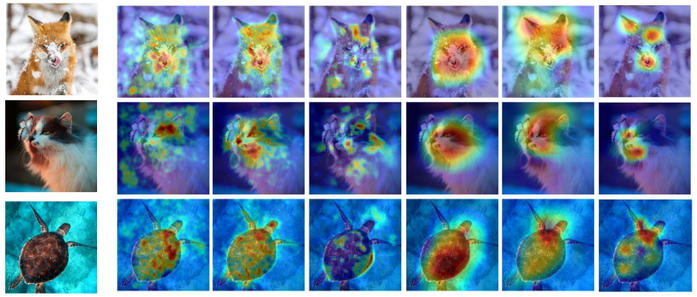
Xplique

A deep learning Explainability Toolbox

Optimized for Tensorflow / Keras ecosystem.

Thomas FEL*, Lucas HERVIER*,
David VIGOUROUX, Antonin POCHE, Justin PLAKOO, Rémi CADENE, Mathieu CHALVIDAL, Julien COLIN, Thibaut BOISSIN, Louis BETHUNE, Agustin PICARD, Claire NICODEME, Laurent GARDES, Grégory FLANDIN, Thomas SERRE

(1) **Attribution Methods** more than **14** black-box / white-box methods



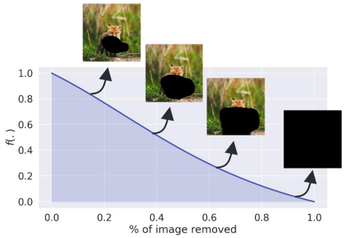

```python
from xplique.attributions import GradCAM

explainer = GradCAM(model)
explanations = explainer(x, y)
```

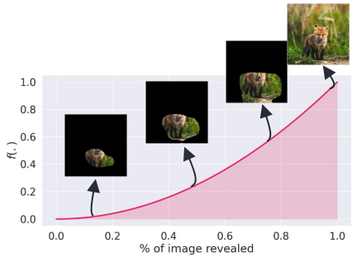
*Pytorch, Sklearn supported for black-box methods

(2) **Metrics** more than **6** attributions metrics each supporting multiple baselines

Deletion (low AUC = better faithfulness)



Insertion* (high AUC = better faithfulness)

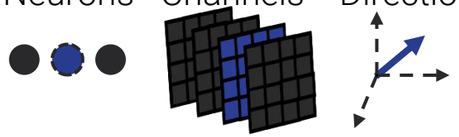



```python
from xplique.metrics import Deletion
from xplique.attributions import GradCAM

metric = Deletion(model, x, y)
explanations = GradCAM(model)(x, y)
score = metric(explanations)
```

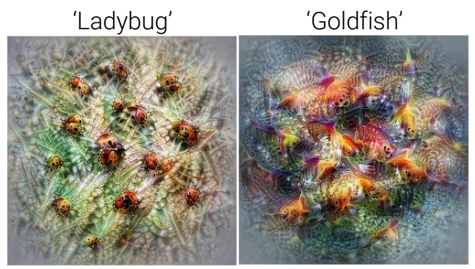
(3) **Feature Visualization**

· Neurons · Channels · Directions



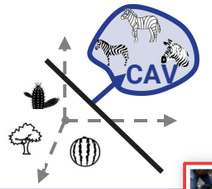
Visualize Neurons, Channels, Vectors in activation space (e.g. CAV) or a mix of them !

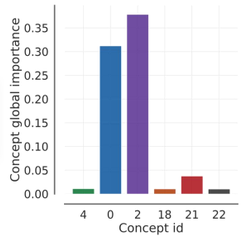
```python
from xplique.feature_visualization import Objective, optimize
obj = Objective.neuron(model, 'logits', 10)
images, obj_name = optimize(obj)
```



(4) **Concept based** concept activation vector, **CRAFT** (new!)

Easily extract and test CAVs:






```python
from xplique.concepts import Cav

extractor = Cav(model, 'mixed3')
concept_vector = extractor(striped_samples,
                           random_samples)
```

github.com/deel-ai/xplique
See also: github.com/deel-ai/deel-lip

Used in:
CRAFT: Concept Activation FacTorization for Explainability
Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis
Don't Lie to Me: Robust & Efficient explainability with Verified Perturbation Analysis
Making Sense of Dependence: Efficient Black-box Explanations Using Dependence Measure
...