iNNvestigate neural networks!
Maximilian Alber et. al.

Tomasz Kanas

19 maja 2021

Po co wyjaśniać sieci neuronowe?

Z tych samych powodów co inne techniki ML, ale też:

Aby lepiej zrozumieć jak działają!

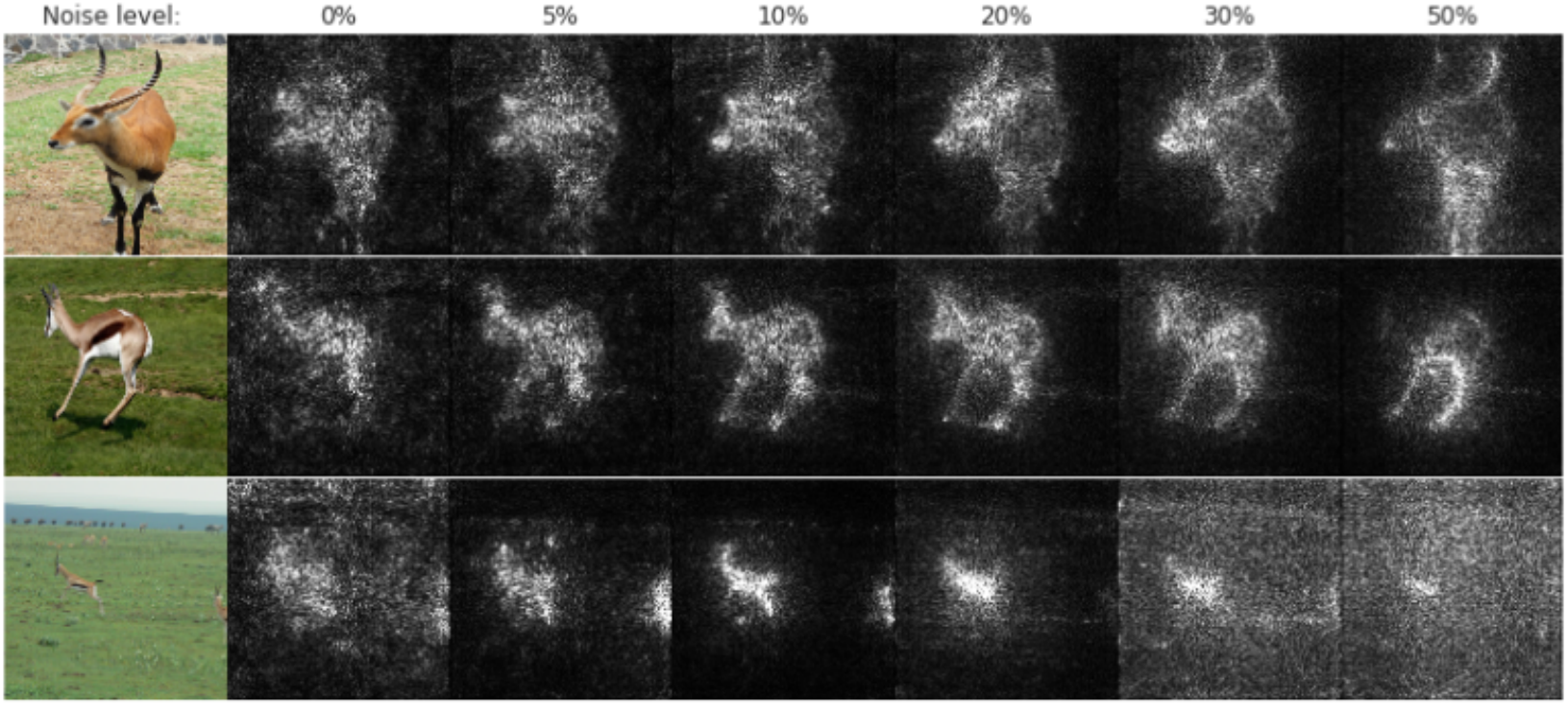
- Porównywanie różnych architektur
- Dobieranie architektury do zadania
- Dobór technik optymalizacyjnych i hiper-parametrów

Plan prezentacji

Wizualizacja ważnych pixeli

Ważność pixela = gradient predykcji względem tego pixela

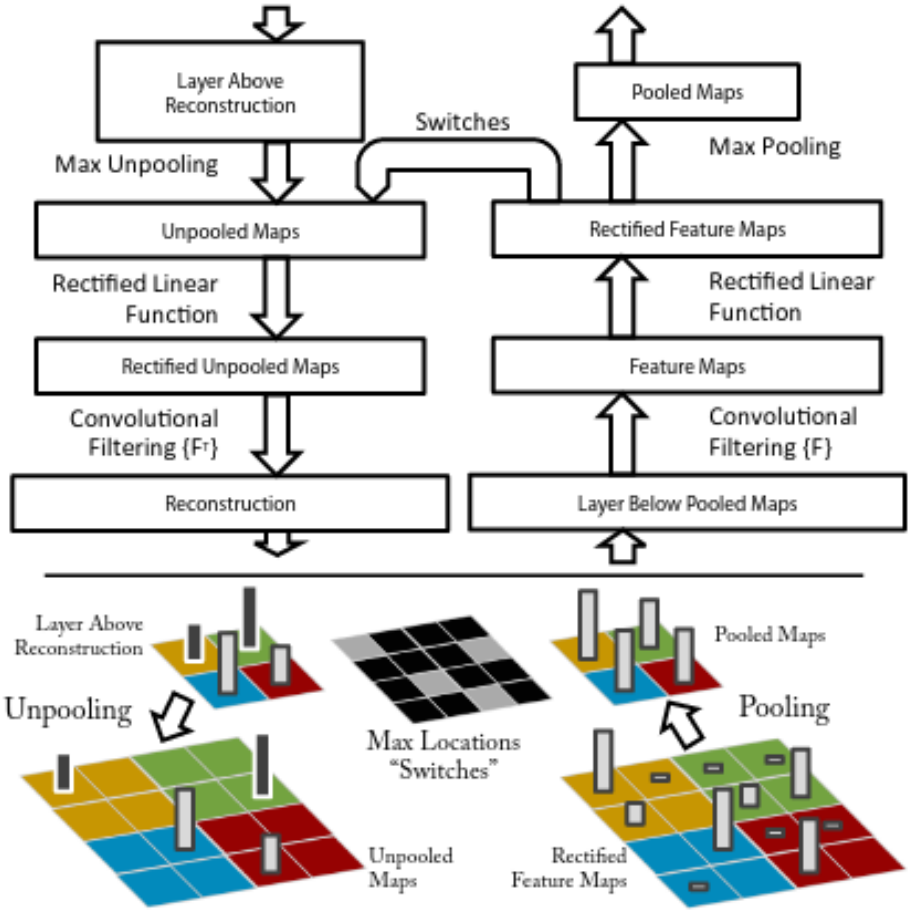
SmoothGrad: uśredniona ważność z kilku ewaluacji obrazka z dodanym szumem losowym.



Szum gaussowski $N(0, \sigma^2)$, średnia z 50 próbek.

Dekonwolucja

Dla wyniku danej warstwy „cofamy” operacje które sieć wykonała i otrzymujemy obrazek cechy którą ta warstwa wykryła.

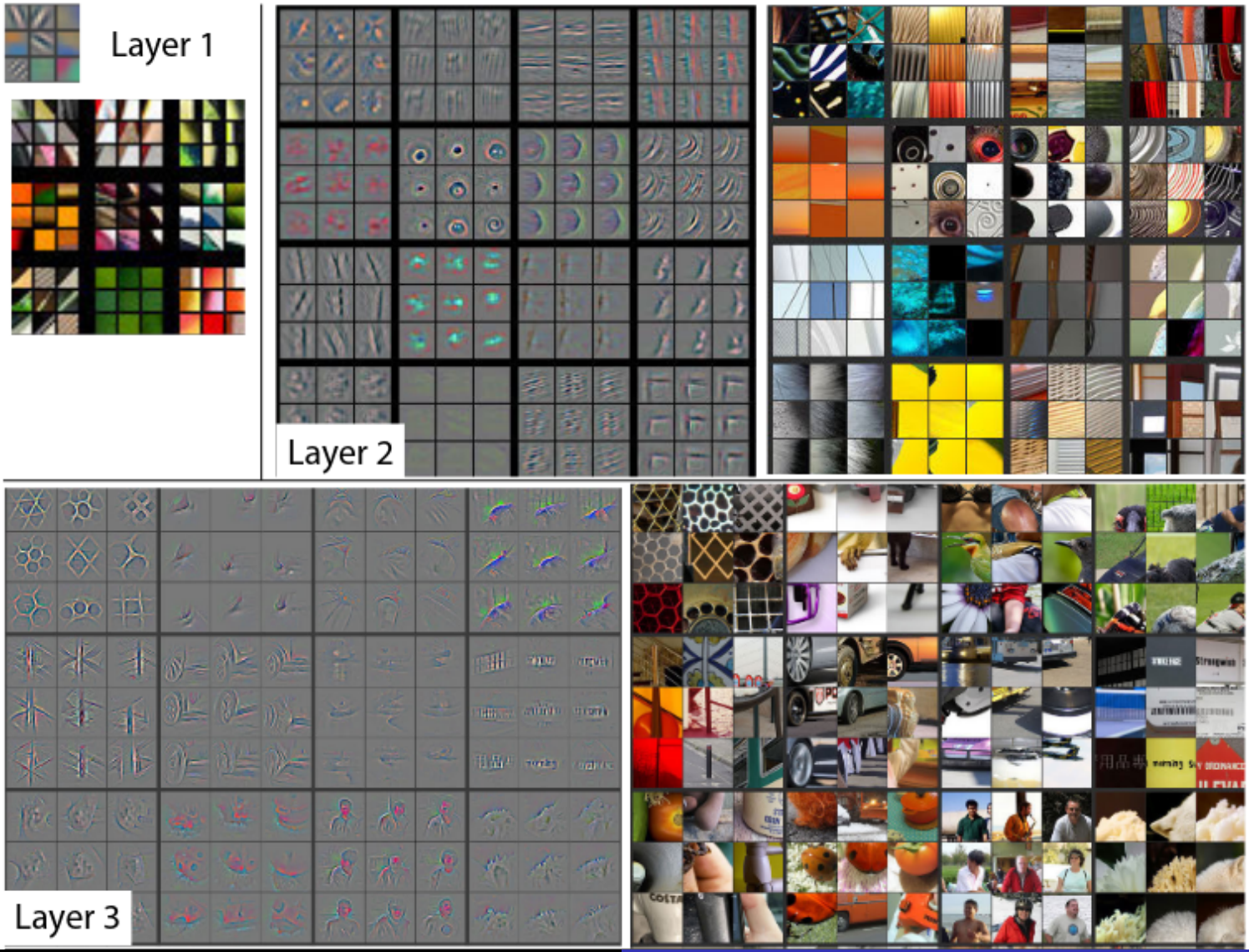


Inne algorytmy:

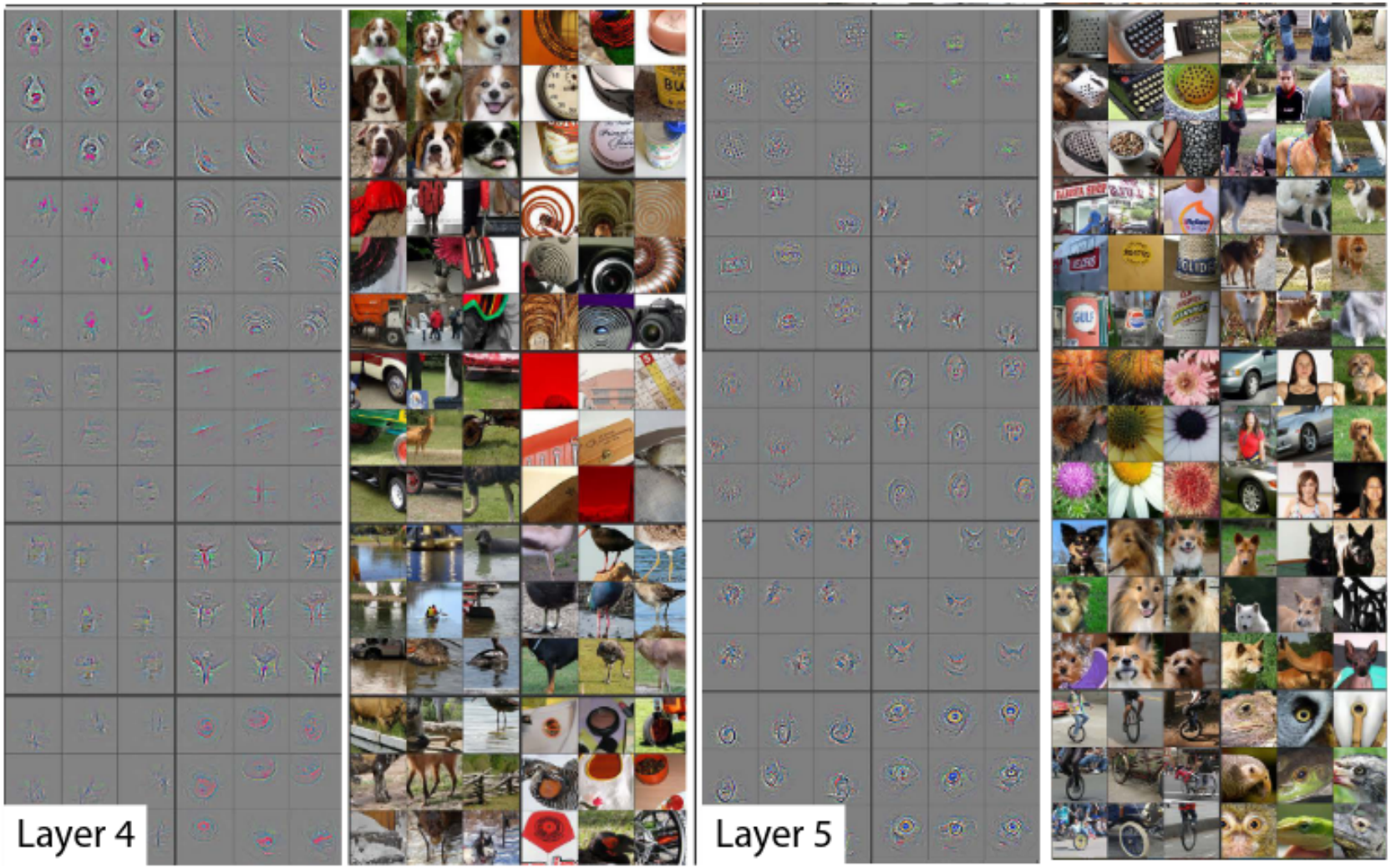
Guided BackProp

PatternNet

DeConvNet: wyniki



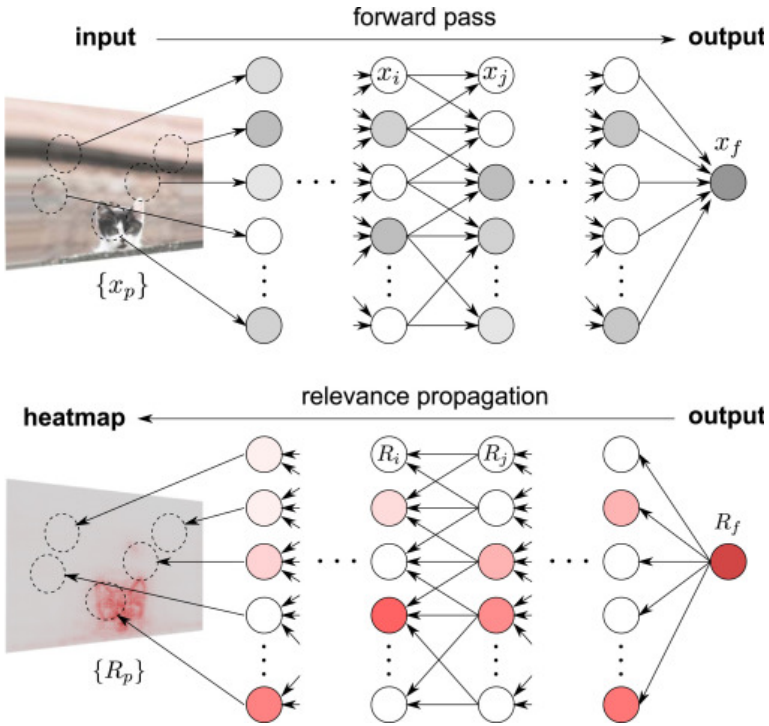
DeConvNet: wyniki



Dekompozycja modelu

Jeśli sieć neuronowa oblicza funkcję $f(x)$ gdzie $x \in \mathbb{R}^d$ to dane wejściowe, to dekompozycja $R(x) \in \mathbb{R}^d$ powinna spełniać

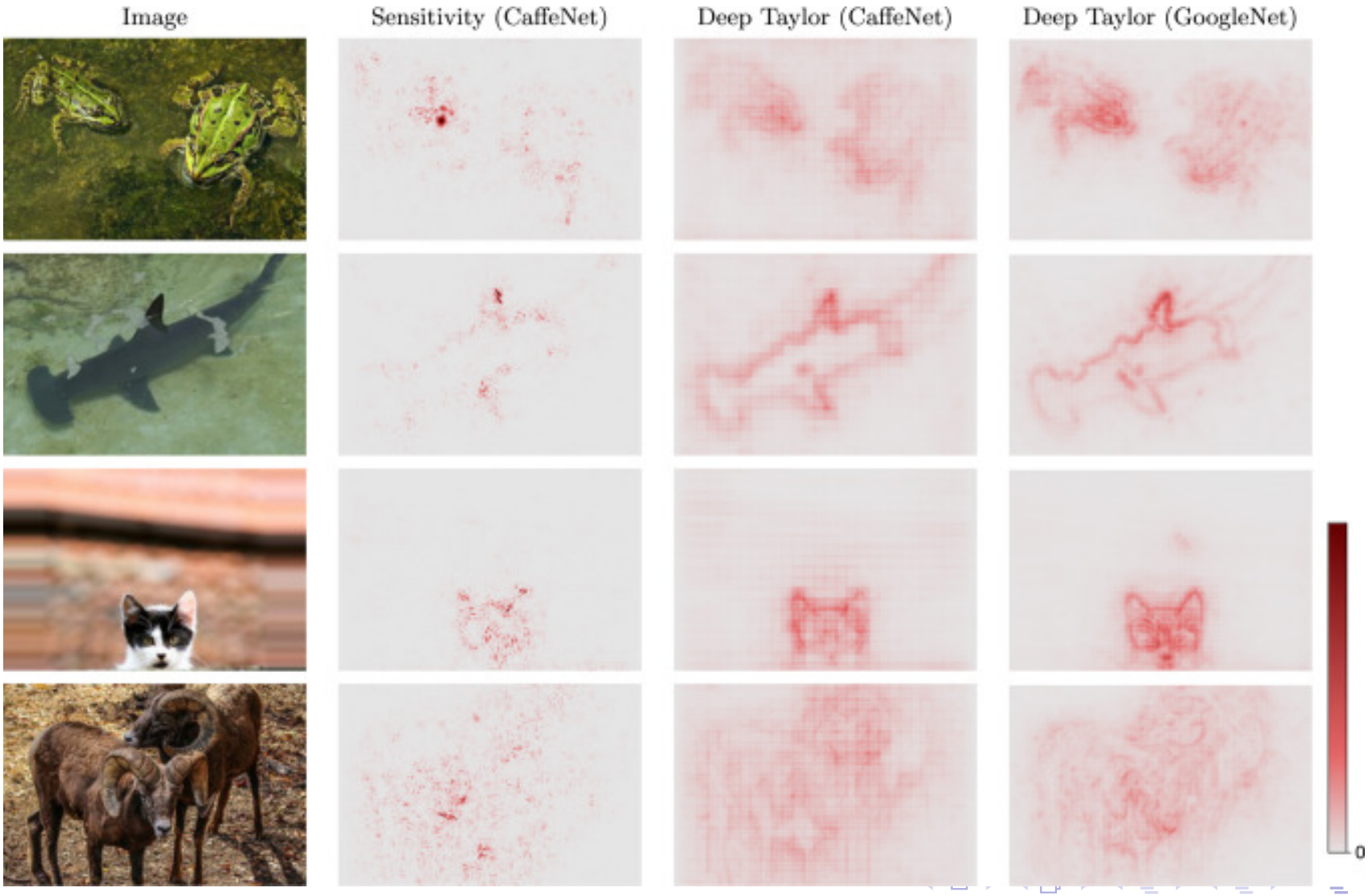
$$\forall_x f(x) = \sum_p R_p(x), \quad \forall_{x,p} R_p(x) \geq 0$$



Idea:

1. Dla każdego neuronu oblicz bliską wartość wejścia dla której wynik jest 0.
2. Przybliż atrybucję neuronu ze wzoru Taylora.
3. Propaguj atrybucje wstecz.

DeepTaylor: wyniki



Alternatywne metody dekompozycji

- Input * Gradient
- PatternAttribution: szuka miejsc zerowych zgodnie z kierunkiem sygnału każdego neuronu.
- LRP: Przydziela rekurencyjnie atrybucję proporcjonalną do wkładu neuronu w wynik
- IntegratedGradients: całkuje gradient wzdłuż ścieżki z wejścia do wyjścia
- DeepLIFT: oblicza bakpropagację atrybucji na podstawie różnicy wyniku neuronu do wyniku „referencyjnego”.

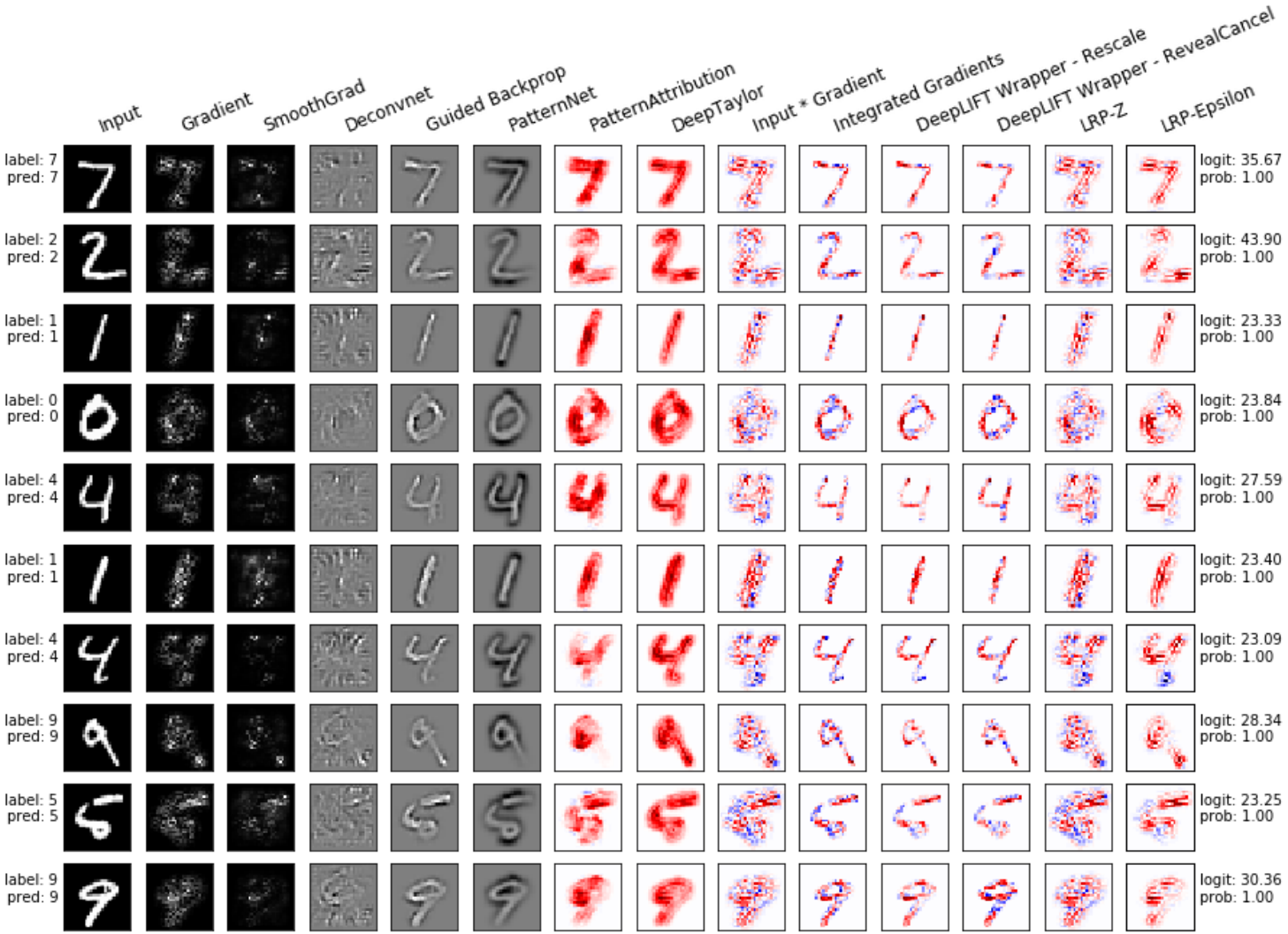
iNNvestigate

Biblioteka pythonowa do porównywania różnych metod wyjaśniania sieci neuronowych.
Umożliwia:

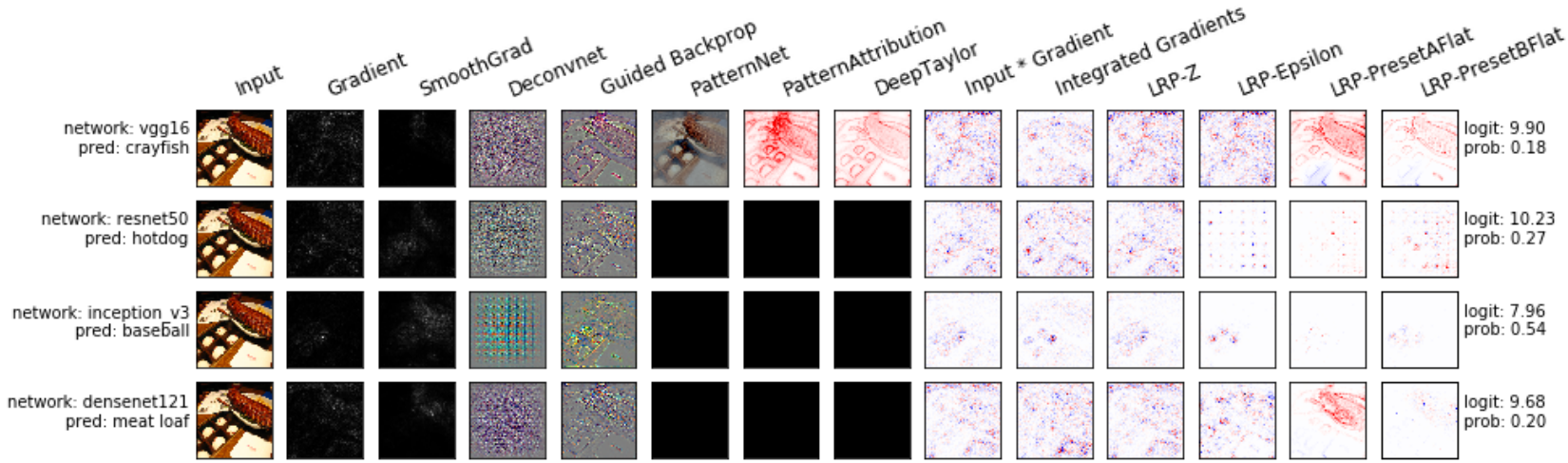
- Trenowanie: niektóre (ie. PatternNet, PatternAttribution) metody wyjaśnialności zależą od rozkładu danych.
- Ilościowa ewaluacja: Sprawdza sensowność wyjaśnień przez perturbowanie regionów mających duży wpływ na wynik (i oczekiwanie innego wyniku).
- Modularność: łatwo dodawać nowe metody

Posiada implementacje omówionych metod.

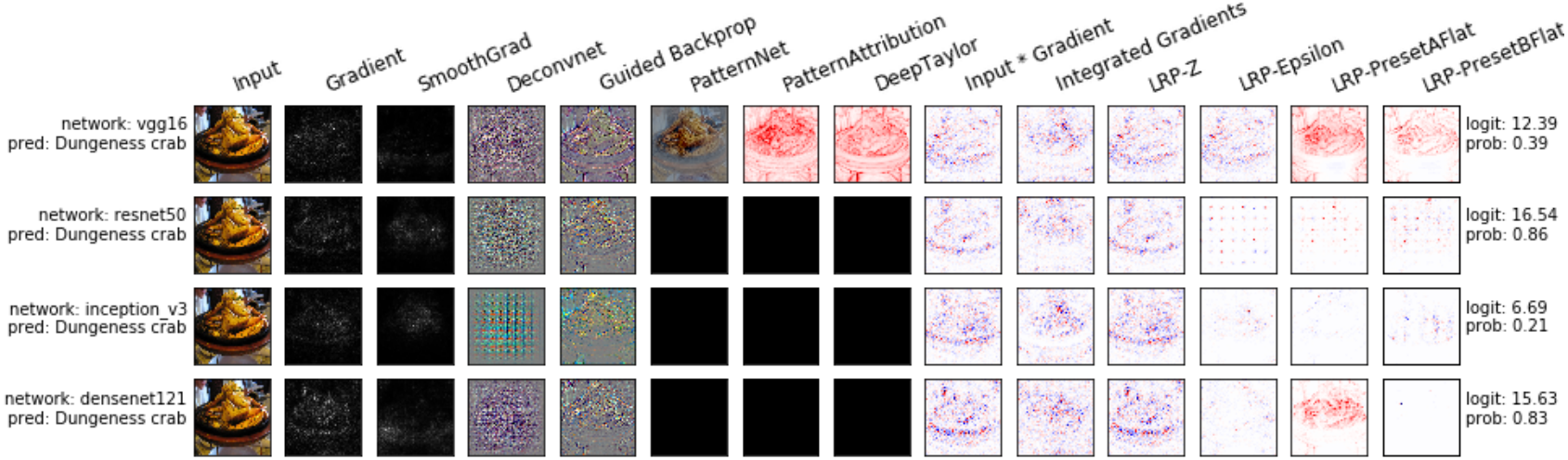
iNNvestigate: MINIST



iNNvestigate: ImageNet



Poprawna odpowiedź: baseball



Dziękuję za uwagę!