Metody interpretacji i rozumienia decyzji podejmowanych przez głębokie sieci neuronowe

Methods for interpreting and understanding deep neural networks
Author: Grégoire Montavon,Wojciech Samek,Klaus-Robert Müller
Publication: Digital Signal Processing
Publisher: Elsevier
Date: February 2018

https://www.sciencedirect.com/science/article/pii/S1051200417302385?via%3Dihub#fg0010

Motywacja

- aby się upewnić, że problem jest prawidłowo reprezentowany i model nie skupia się/wykorzystuje nieistotne szczegóły danych => dodatkowa walidacja modelu

- w medycynie, pojazdach autonomicznych i tam gdzie wymagane jest poleganie na modelu trzeba mieć pewność, że model wykorzystuje właściwe cechy

- aby zdobyć nowy wgląd w skomplikowane systemy w nauce

Interpretacja post-hoc

- ten typ interpretacji zakłada, że mamy gotowy wytrenowany model klasyfikatora i chcemy zrozumieć co ten model przewiduje na podstawie zrozumienia typowych wzorców
- jest to w kontraście do budowania modeli, które w swojej strukturze zawierają element objaśniania => np. modele liniowe w statystce, drzewa decyzyjne

Rozumienie modelu

- chodzi to o rozumienie funkcjonalne, a nie o rozumienie algorytmiki modelu

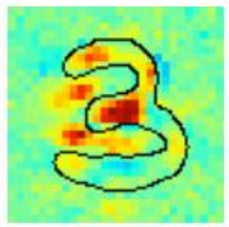
- rozumienie zachowania vs. rozumienie obliczeń wykonywanych wewnątrz czarnej skrzynki

Rozumienie modelu

- chodzi to o rozumienie funkcjonalne, a nie o rozumienie algorytmiki modelu

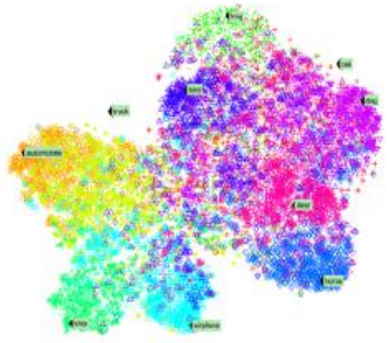
- rozumienie zachowania vs. rozumienie obliczeń wykonywanych wewnątrz czarnej skrzynki

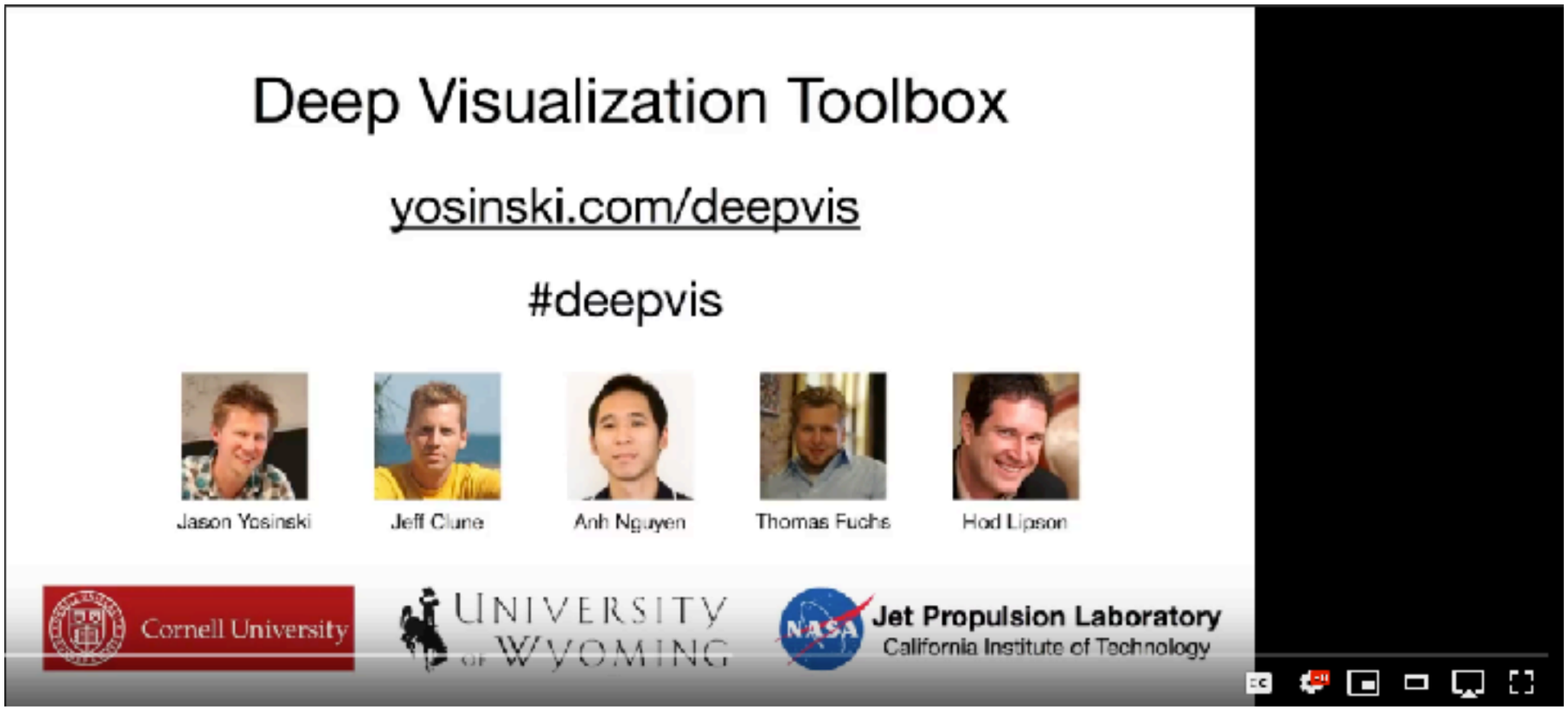
predykcja
dlaczego pewien wzorzec x został zaklasyfikowany jako klasa ω



model
jak wygląda typowy wzorzec należący do danej kategorii wg. modelu

dane
które wymiary danych są najbardziej istotne dla tego problemu?





http://yosinski.com/deepvis

Definicje

- **interpretacja** to mapowanie abstrakcyjnego pojęcia (przewidywanej klasy) do dziedziny, w której *człowiek* może zrozumieć sens tej predykcji. Taką dziedziną mogą być np. obrazy, teksty itp.

- **wyjaśnienie** to kolekcja cech w dziedzinie zrozumiałej dla *człowieka*, które przyczyniły się do danej predykcji.
  Cechom tym można też przypisać pewne wagi.

  - przykładem wyjaśnienia może być mapa kolorująca piksle w zależności od ich wkładu do decyzji

  - w klasyfikacjach związanych z językiem może to być podświetlenie słów i zwrotów przyczyniających się do decyzji

- Ten nacisk na rozumienie i interpretację przez człowieka jest też związany z aspektem prawnym „przypisania odpowiedzialności” i „prawa do wyjaśnienia”

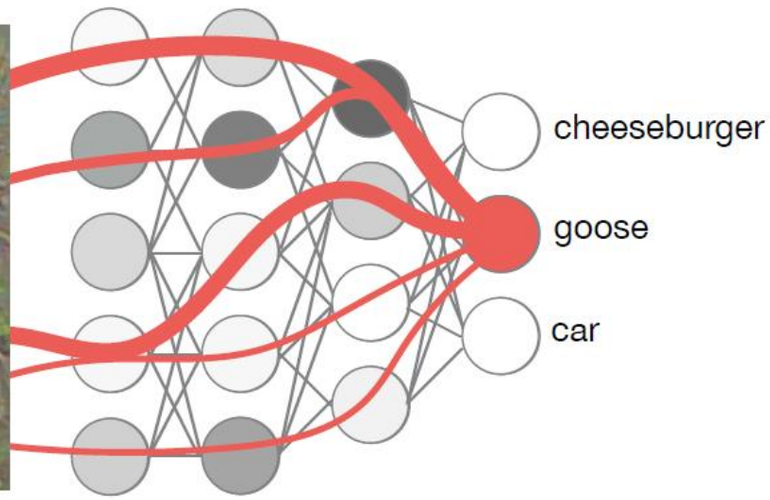
Interpretacja modelu

- znajdź prototypowy przykład dla danej klasy
- znajdź wzorzec maksymalizujący aktywność danego neuronu

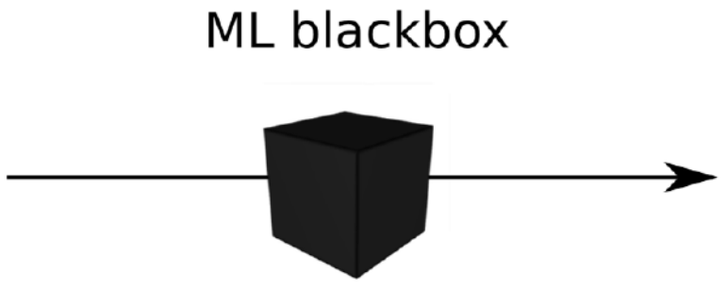
Wyjaśnianie decyzji

- dlaczego model wykonał taką klasyfikacje
- weryfikacja, że model działa zgodnie z naszą intuicją i rozumieniem problemu

simple regularizer (Simonyan et al. 2013)



$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x) + \lambda \Omega(x)$$



Popularne podejścia do wizualizacji cech

- Dekonwolucja:
  - https://www.matthewzeiler.com/mattzeiler/adaptivedeconvolutional.pdf
  - Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53

- Guided backpropagation:
  - Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplic- ity: the all convolutional net. In: ICLR Workshop (2015) https://arxiv.org/pdf/1412.6806.pdf

Interpretacja sieci głębokiej

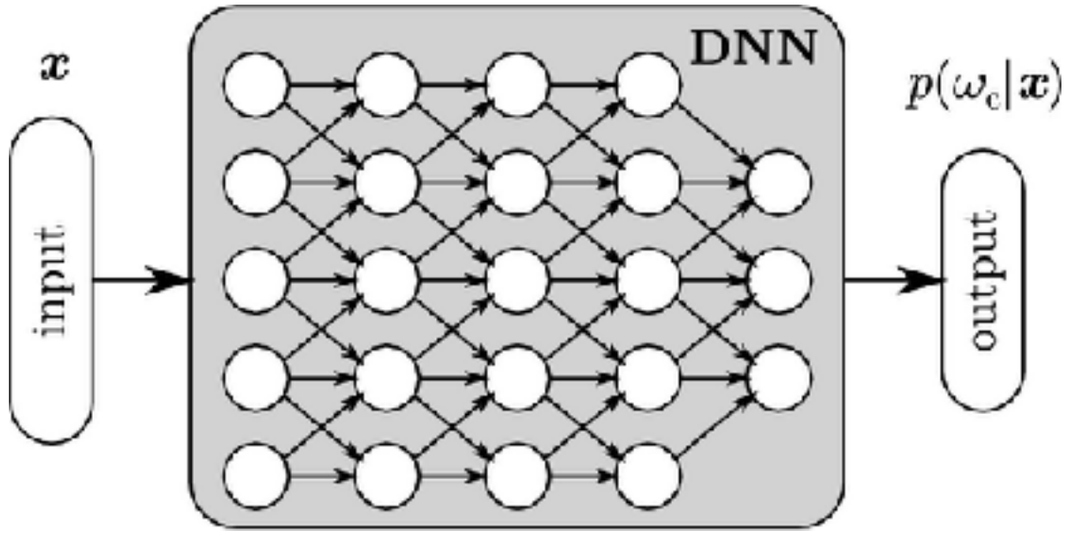
- Neurony w sieci łącznie tworzą skomplikowane nieliniowe mapowanie z przestrzeni cech do przestrzeni klas

- neurony wyjściowe odpowiadają abstrakcyjnym pojęciom

- przestrzeń wejściowa jest interpretowalna (obrazy są interpretowalne dla człowieka)

- skupimy się teraz na tworzeniu prototypu w przestrzeni wejść, który reprezentuje wyuczone abstrakcyjne pojęcie

    - stworzymy go w ramach podejścia maksymalizacji aktywności



Maksymalizacja aktywności

- Poszukiwanie takiego wzorca wejściowego, który maksymalizuje wyjście dla pewnej klasy

$$\boldsymbol{x}^{\star} = \max_{\boldsymbol{x}} \log p(\omega_c | \boldsymbol{x}) - \lambda \|\boldsymbol{x}\|^2$$

- $(\omega_c)_c$ zbiór klas

- $p(\omega_c | \boldsymbol{x})$ - prawdopodobieństwo przynależności do klasy zwracane przez neuron wyjściowy.

- prawdopodobieństwa klas można maksymalizować metodami gradientowymi

- ta metoda zastosowana od obrazów zwraca obrazki w większości szare z kilkoma najważniejszymi pikslami lub krawędziami

- prototypy choć maksymalizują wyjście to nie wyglądają dla człowieka naturalnie

Ulepszanie MA

- zamiast regularyzacji ℓ_2 można użyć innych, zawierających pewną wiedzę o danych, np. ich rozkład prawdopodobieństwa

$$\max_{\boldsymbol{x}} \log p(\omega_c \,|\, \boldsymbol{x}) + \log p(\boldsymbol{x})$$

- ta regularyzcja prowadzi przez regułę Byesa do

$$\max_{x} \log p(\omega_c \,|\, x) p(x) = \max_{x} \log p(x \,|\, \omega_c)$$

  czyli do wzorca w ciągu uczącym, który jest najbardziej charakterystyczny dla swojej klasy

- opublikowanych pomysłów na regularyzację jest więcej

Algorytm MA w przestrzeni abstrakcyjnych kodów

- modele generatywne - metoda ucznia bez nauczyciela. Modele te nie dają wprost funkcji gęstości $p(x)$, ale potrafią z niej próbkować:

  - pobierają próbkę z jakiegoś prostego rozkładu $q(z) \sim \mathcal{N}(0, I)$ zdefiniowanego w abstrakcyjnej przestrzeni kodów $\mathcal{Z}$

  - Stosują funkcję dekodującą $g : \mathcal{Z} \to \mathcal{X}$, która mapuje do oryginalnej przestrzeni wejść $\mathcal{X}$

- Przykładem takich modeli są GAN generative adversarial network

Algorytm MA w przestrzeni abstrakcyjnych kodów

forward pass

backward pass

Code h

Image $G(h)$

Label

candle

banana

convertible

Generator network (prior)

DNN being visualized

A. Nguyen et al. w W. Samek et al. (Eds.): Explainable AI, LNAI 11700, pp. 55–76, 2019.

Algorytm MA w przestrzeni abstrakcyjnych kodów

- Nguyen zaproponował wbudowanie takiego generatora wprost do algorytmu MA; jego problem optymalizacyjny jest następujący:

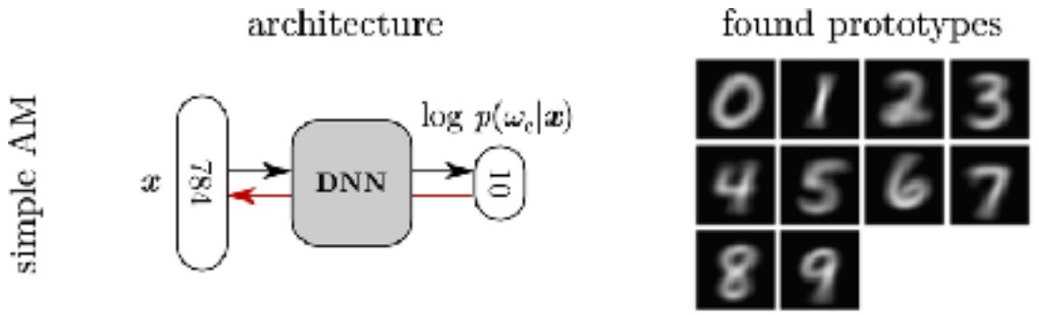
$$\max_{z \in \mathcal{Z}} \log p(\omega_c \mid g(z)) - \lambda \|z\|^2$$

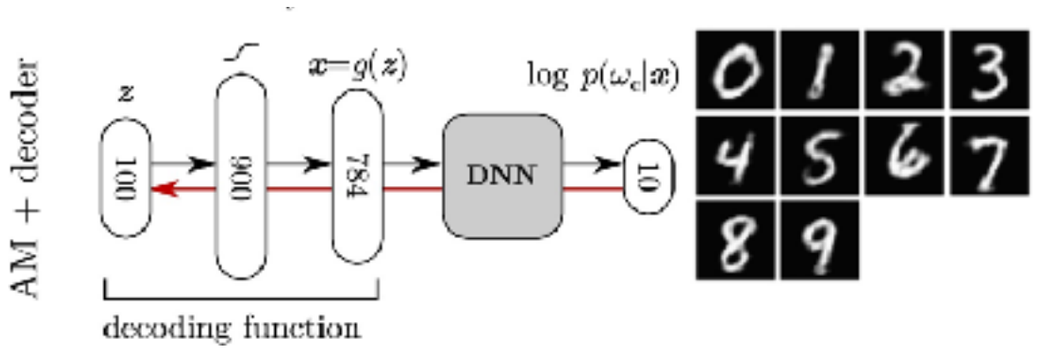
- po znalezieniu optymalnego z^* prototyp dla klasy ω_c jest znajdowany jako $x^* = g(z^*)$

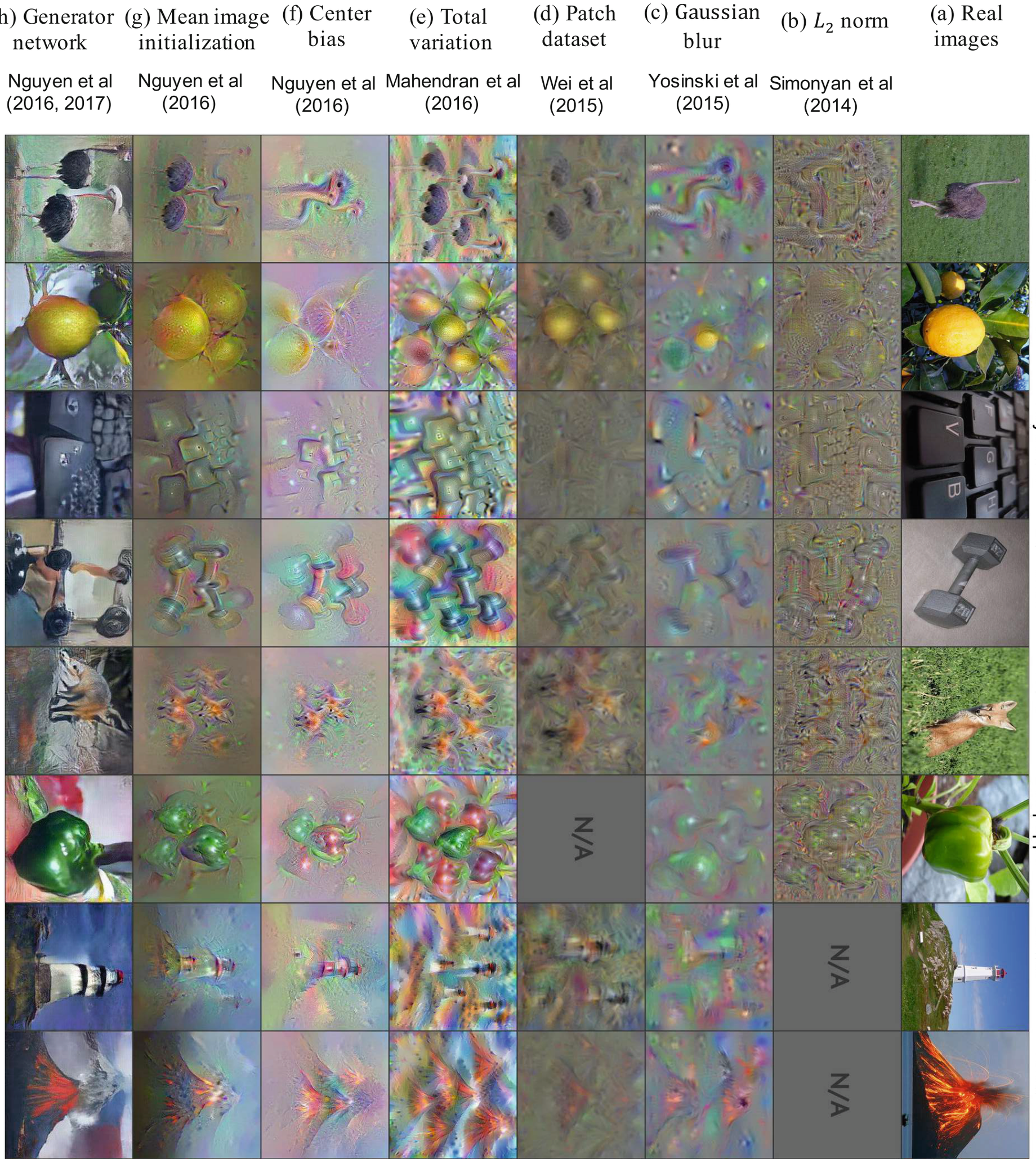
- dla normalnego rozkładu kodów $q(z)$ człon regularyzacyjny jest równoważny $\log q(z)$, czyli faworyzuje kody o wysokim prawdopodobieństwie

Interpretacja klasyfikacji zbioru MNIST

- prosta regulryzacja normą l_2
- z modelem generatywnym w postaci dwuwarstwowego dekodera, regularyzacja normą l_2 , przy czym w funkcji celu wyrażenie było postaci $\lambda \|z - \bar{z}\|^2$, $\bar{z}$ oznacza średni kod dla klasy ω_c



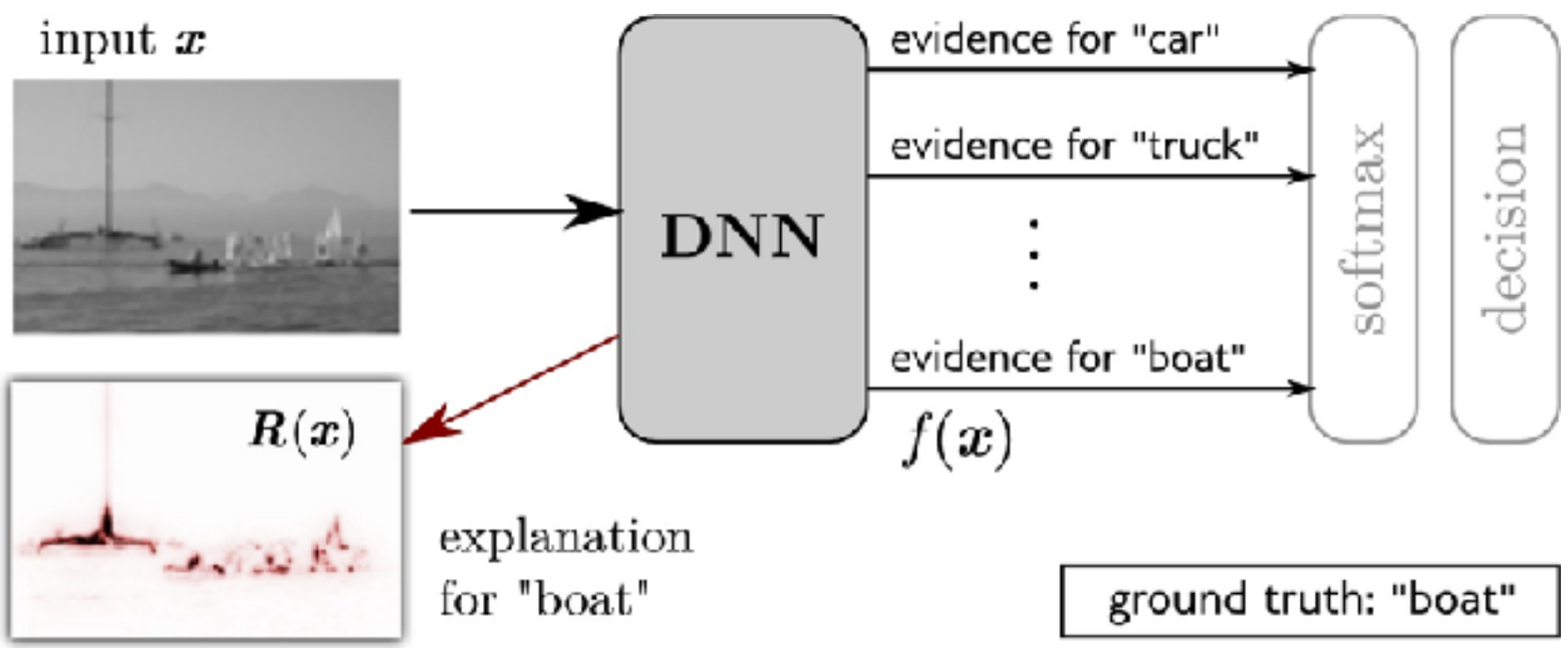




W. Samek et al. (Eds.): Explainable AI, LNAI 11700, pp. 55–76, 2019.
https://doi.org/10.1007/978-3-030-28954-6_4

Bardziej skomplikowane klasy - wyjaśnianie DNN

- Pytanie: „Jakie cechy wzorca x powodują, że jest on dobrym reprezentantem klasy ω_c?
- głęboka sieć neuronowa produkuje wyjście $f(x)$
- patrzymy na wzorzec x jako na zbiór cech $(x_i)_{i=1}^{d}$
- dla każdej cechy przypisujemy rangę R_i , która mówi jak istotna jest cecha x_i dla wyjaśnienia $f(x)$



Analiza czułości

- $R_i(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_i}\right)^2$,

  gradient jest obliczy w punkcie x.

- w tym sensie najbardziej istotne cechy to to, które powodują największe zmiany funkcji $f(x)$

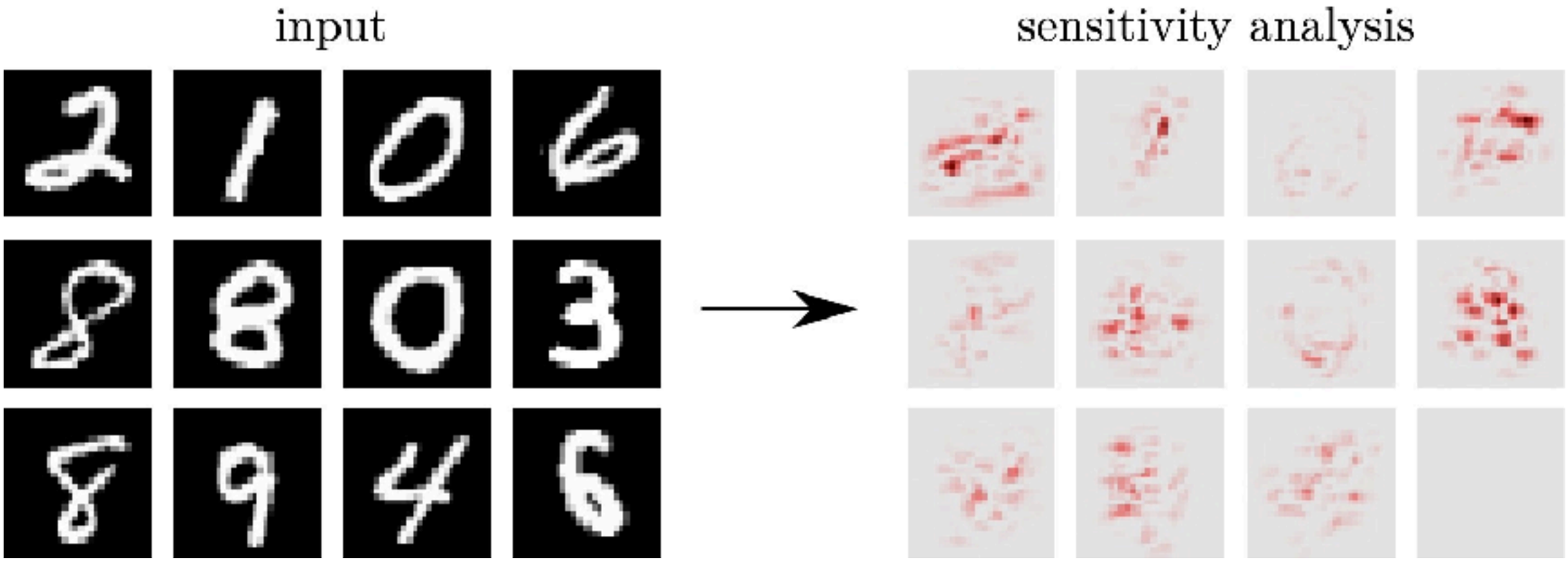
- zaletą jest to, że gradienty można wyliczać w ramach algorytmu wstecznej propagacji

- ta analiza skupia się na zmianach funkcji, a nie na jej wartościach

  $\sum_{i=1}^{d} R_i(\boldsymbol{x}) = \|\nabla f(\boldsymbol{x})\|^2$

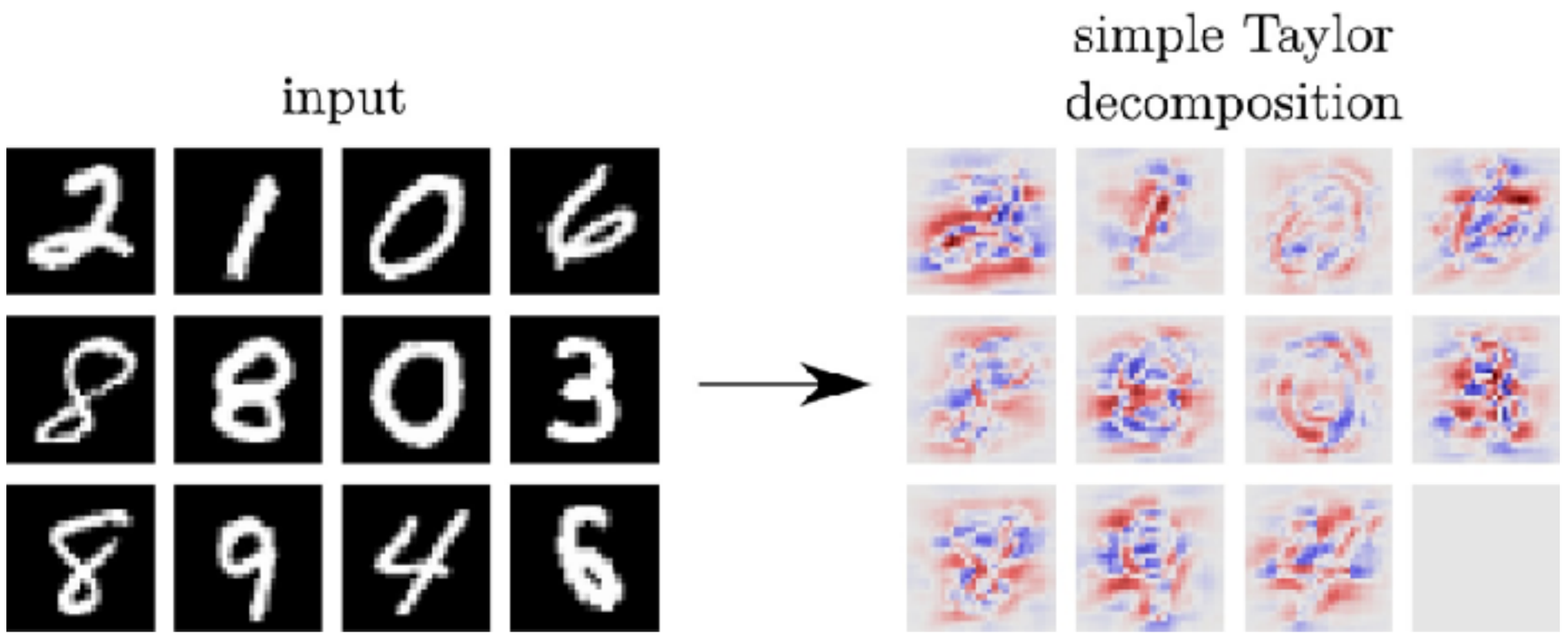
- mapa termiczna wskazuje, które piksele powodują, że cyfra należy do klasy docelowej *bardziej / mniej*

- nie wskazuje , co sprawia, że cyfra należy do tej klasy.



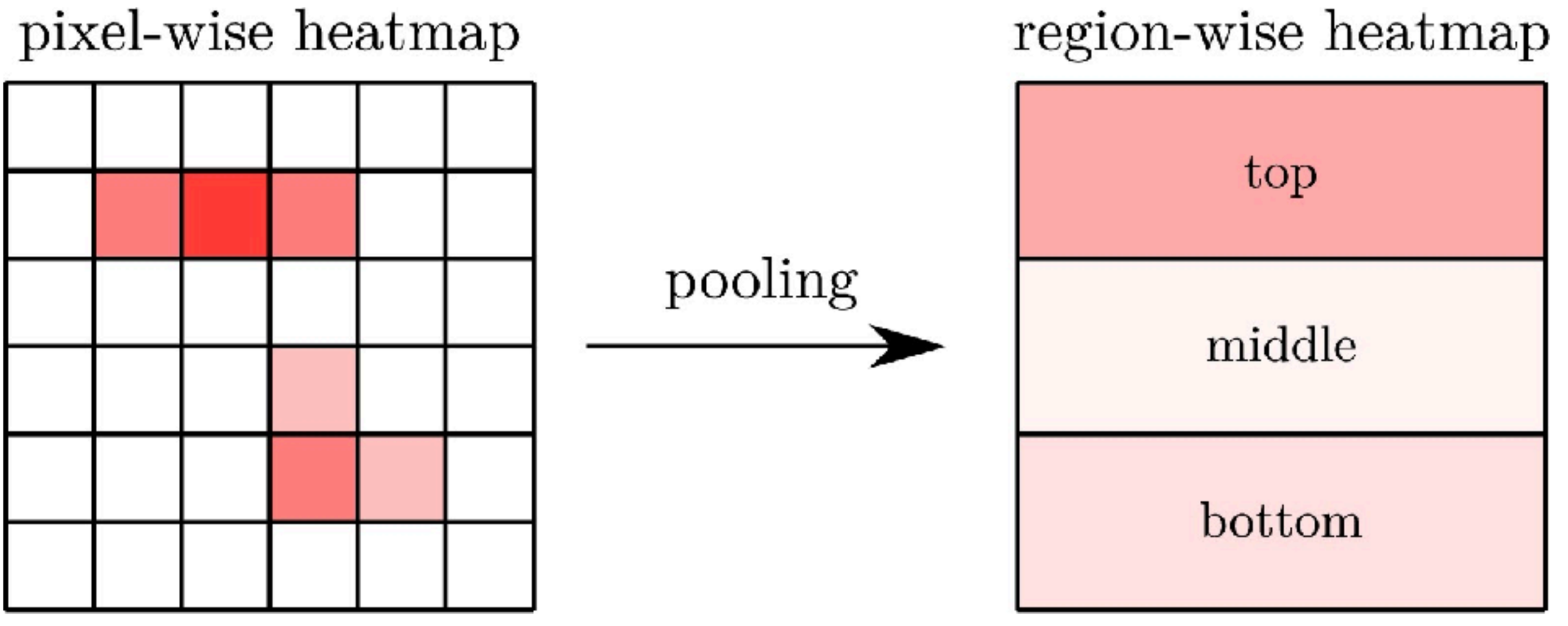
Proste rozwinięcie Taylora

- Rozwinięcie Taylora jest w rozważanym kontekście zapisaniem $f(x)$ jako sumy rang istotności
- rangi są otrzymywane jako rozwinięcia $f(x)$ pierwszego rzędu wokół puntu $\tilde{x}$ dla którego $f(\tilde{x}) = 0$, zatem
  $f(\boldsymbol{x}) = \sum_{i=1}^{d} R_i(\boldsymbol{x}) + O(\boldsymbol{x}\boldsymbol{x}^\top)$
  gdzie
  $R_i(\boldsymbol{x}) = \left.\frac{\partial f}{\partial x_i}\right|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}} \cdot (x_i - \tilde{x}_i)$



Grupowanie cech

- Jeśli funkcję $f(x)$ przybliżamy przez sumę cech to składniki tej sumy można pogrupować
- $R_{\mathcal{J}}(\boldsymbol{x}) = \sum_{i \in \mathcal{J}} R_i(\boldsymbol{x})$
- $f(\boldsymbol{x}) = \sum_{\mathcal{J}} R_{\mathcal{J}}(\boldsymbol{x})$

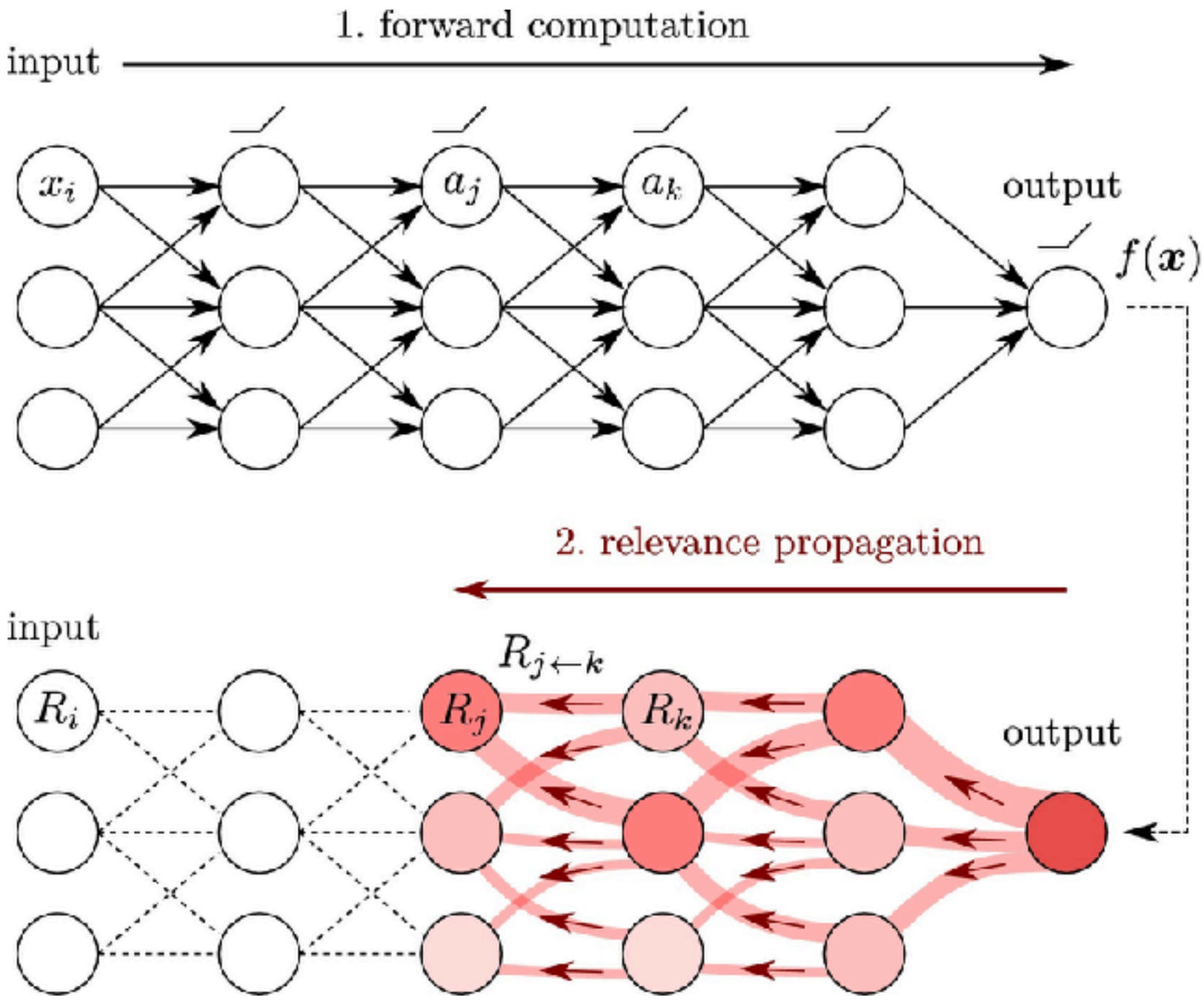


Techniki propagacji wstecznej

- ogólny pomysł polega na wykorzystaniu struktury grafu tworzącego sieć:

  Zaczynamy od wyjścia sieci. Następnie poruszamy się po grafie w odwrotnym kierunku, stopniowo mapując prognozę na niższe warstwy. Procedura kończy się po osiągnięciu wejścia sieci.

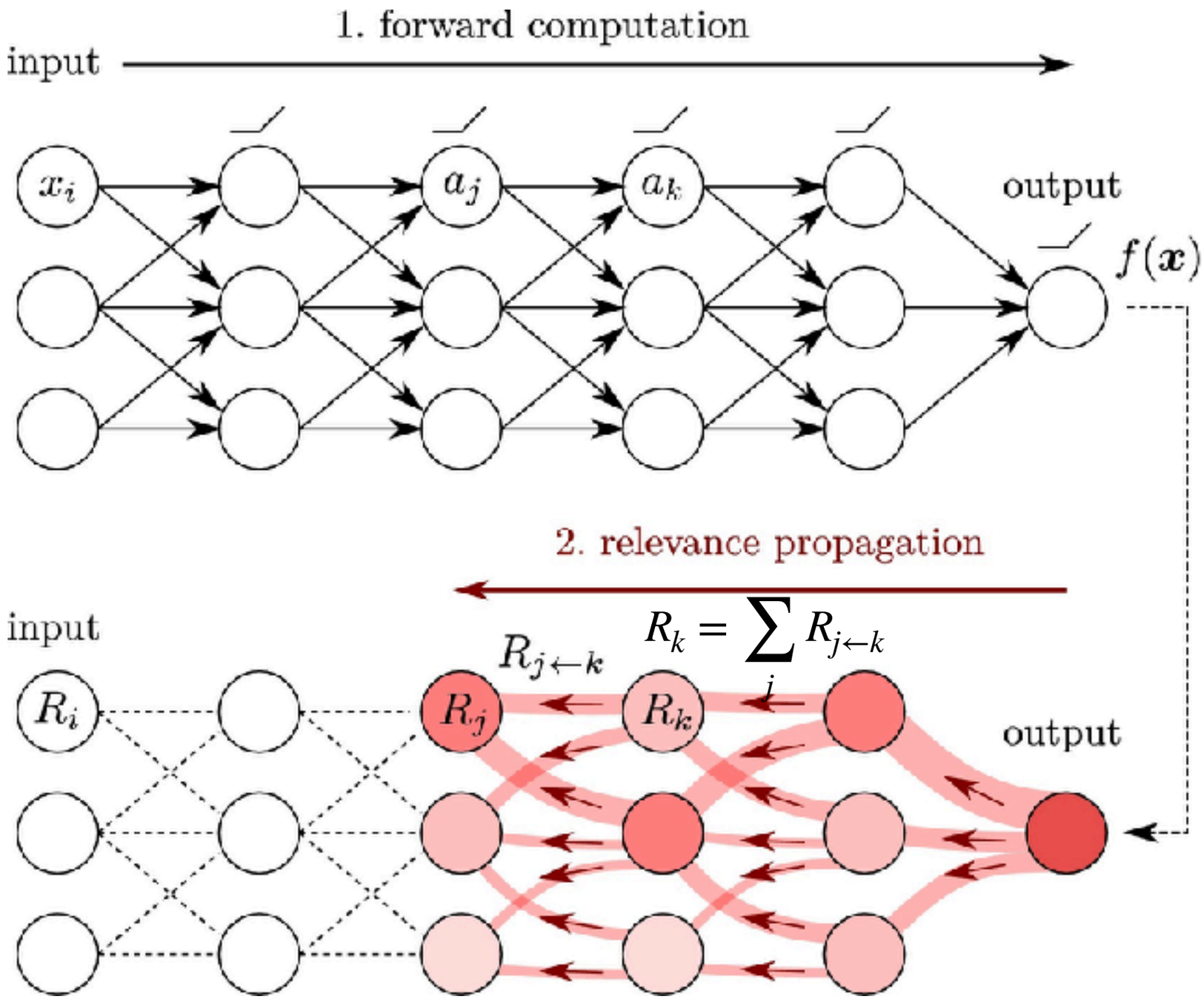
Layer-wise relevance propagation (LRP) -warstwowa propagacja istotności (Bach et al. 2015)



$$f(x) = \sum_i R_i$$

- **Każdy neuron dostaje udział w istotności proporcjonalnie do swojej aktywacji i do siły połączenia**
- **Obowiązuje zasada zachowania istotności**

Layer-wise relevance propagation (LRP) -warstwowa propagacja istotności (Bach et al. 2015)

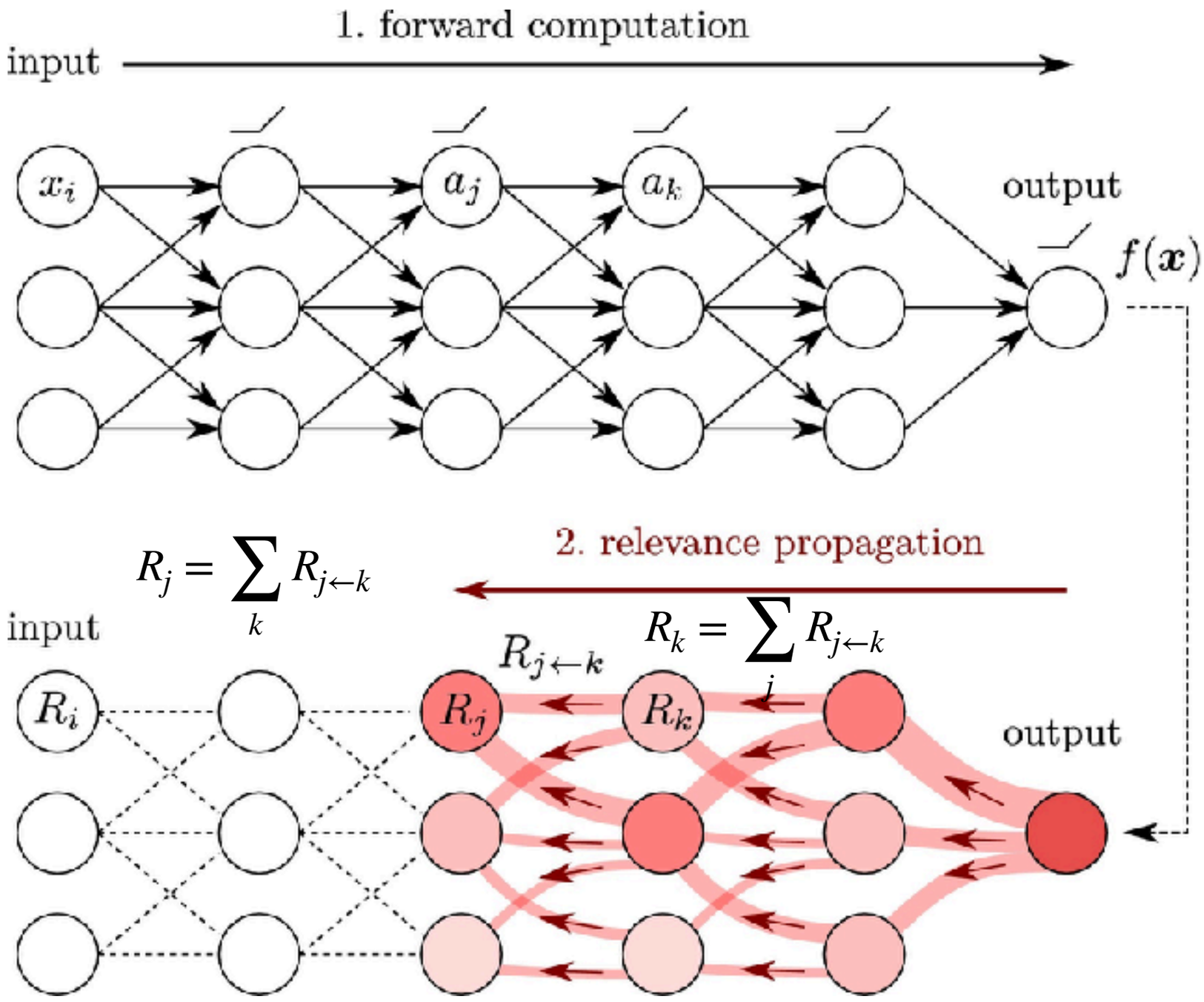


$$R_k = \sum_j R_{j\leftarrow k}$$

$$f(x) = \sum_i R_i$$

- **Każdy neuron dostaje udział w istotności proporcjonalnie do swojej aktywacji i do siły połączenia**
- **Obowiązuje zasada zachowania istotności**

Layer-wise relevance propagation (LRP) -warstwowa propagacja istotności (Bach et al. 2015)



- **Każdy neuron dostaje udział w istotności proporcjonalnie do swojej aktywacji i do siły połączenia**
- **Obowiązuje zasada zachowania istotności**

Layer-wise relevance propagation (LRP) -warstwowa propagacja istotności (Bach et al. 2015)

input x

Classifier

Rooster

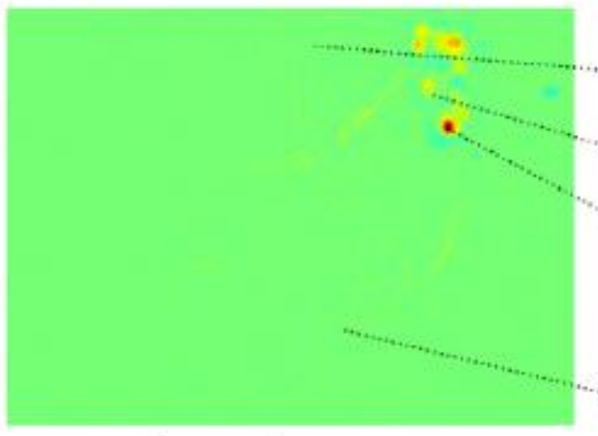
prediction $f(x)$

Explain prediction

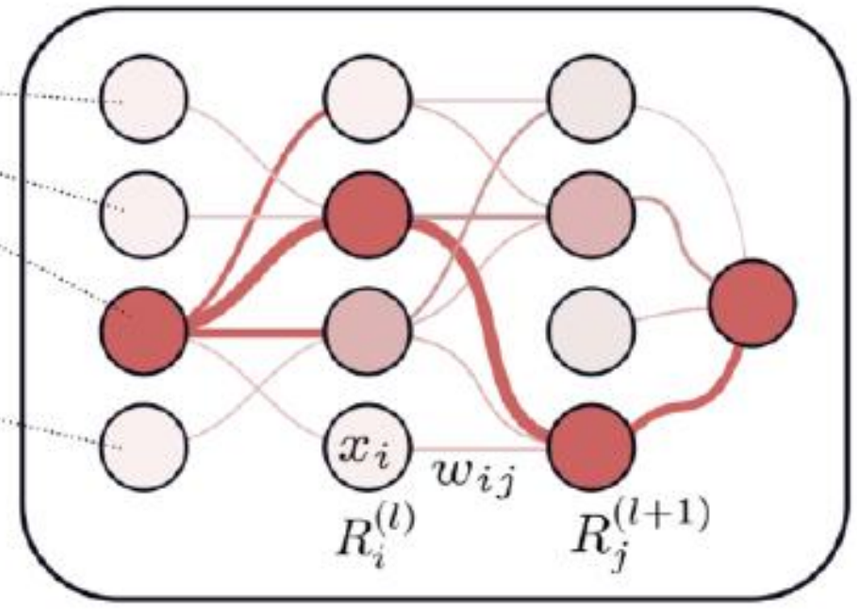
(how much each pixel contributes to prediction)

Idea: Decompose function

$$\sum_i R_i = f(x)$$



heatmap



"every neuron gets it's share of relevance depending on activation and strength of connection."

redistribute $f(x)$

Theoretical interpretation
Deep Taylor Decomposition
(Montavon et al., 2017)

Zasady propagowania istotności

- niech aktywność neuronu będzie

$$a_k = \sigma\left(\sum_j a_j w_{jk} + b_k\right)$$

- zasada $\alpha\beta$:

$$R_j = \sum_k \left(\alpha \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} - \beta \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-}\right) R_k,$$

- z dodatkowym warunkiem: $\alpha - \beta = 1$ i $\beta > 0$

- regułę tą można przepisać jako:

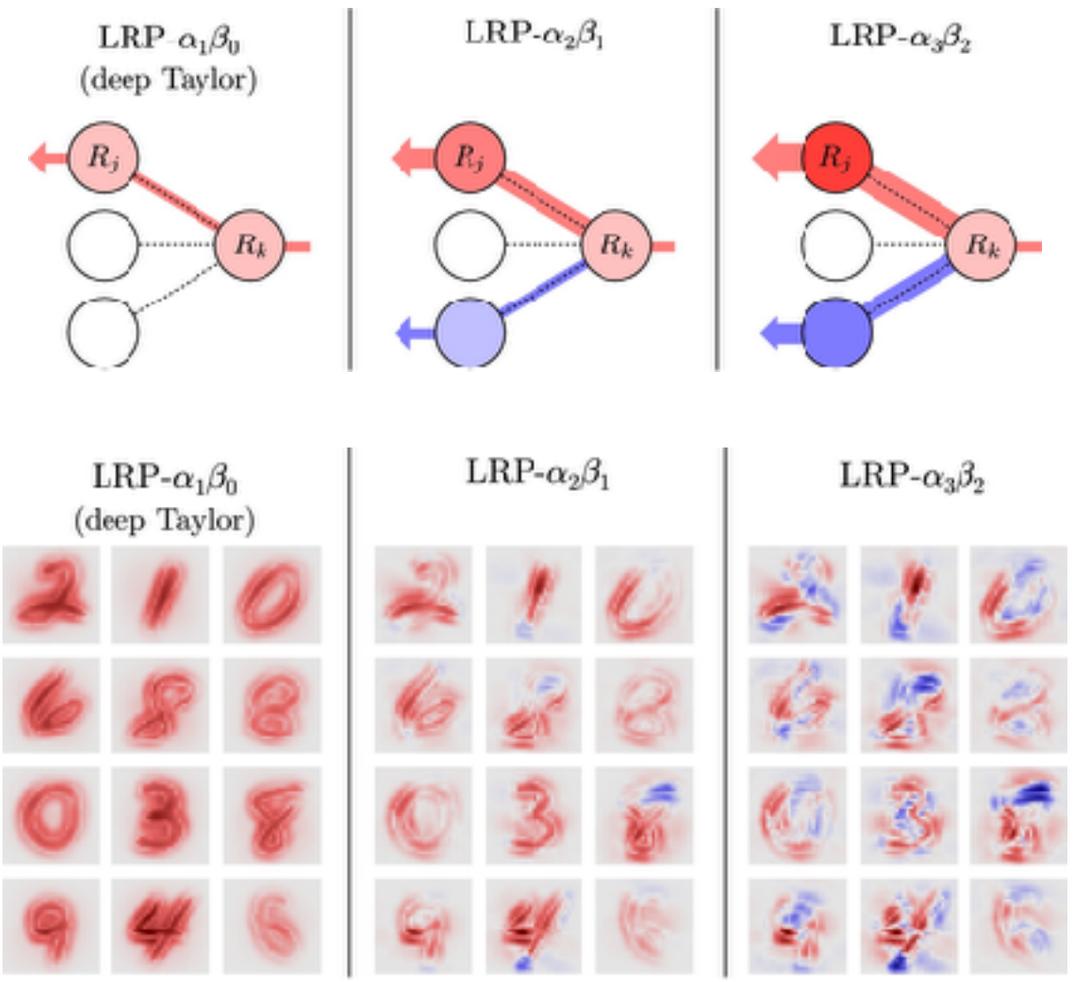
$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k^\wedge + \sum_k \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} R_k^\vee,$$

- tu istotność pozytywna $R_k^\wedge = \alpha R_k$
  i istotność negatywna „przeciw-istotność”
  $R_k^\vee = -\beta R_k$

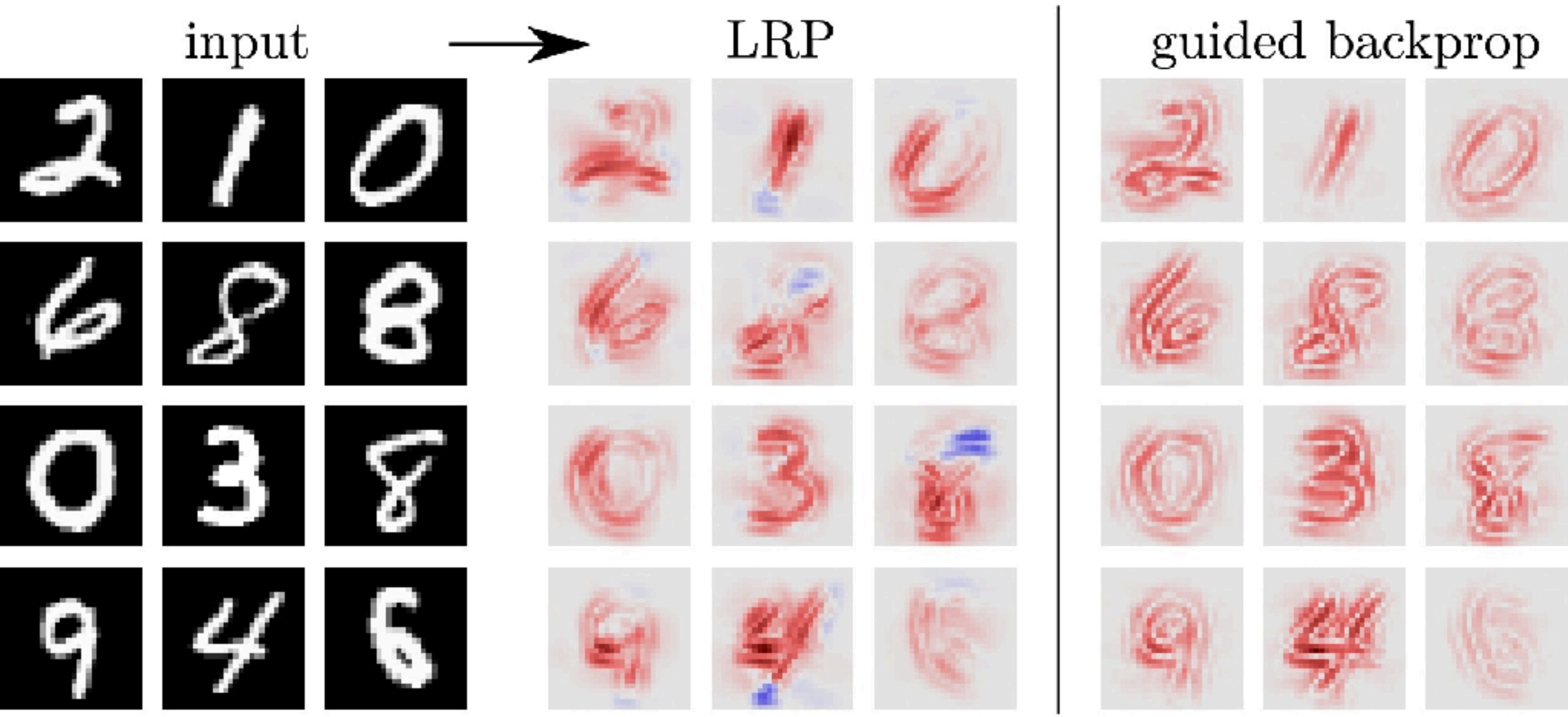
- parametry α, β można wybierać różnie

- suma istotności jest stała

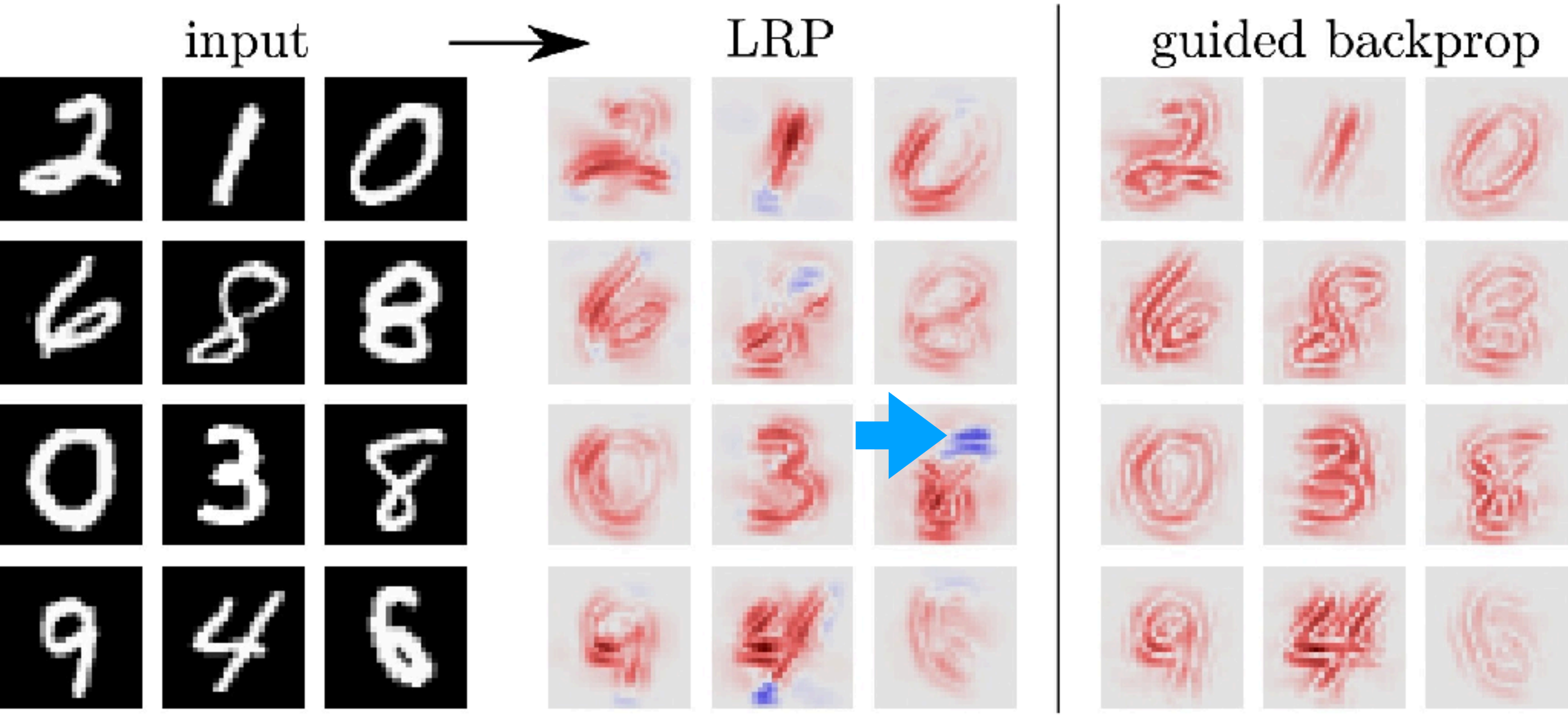
$$(w_{jk})_j = (1, 0, -1)$$



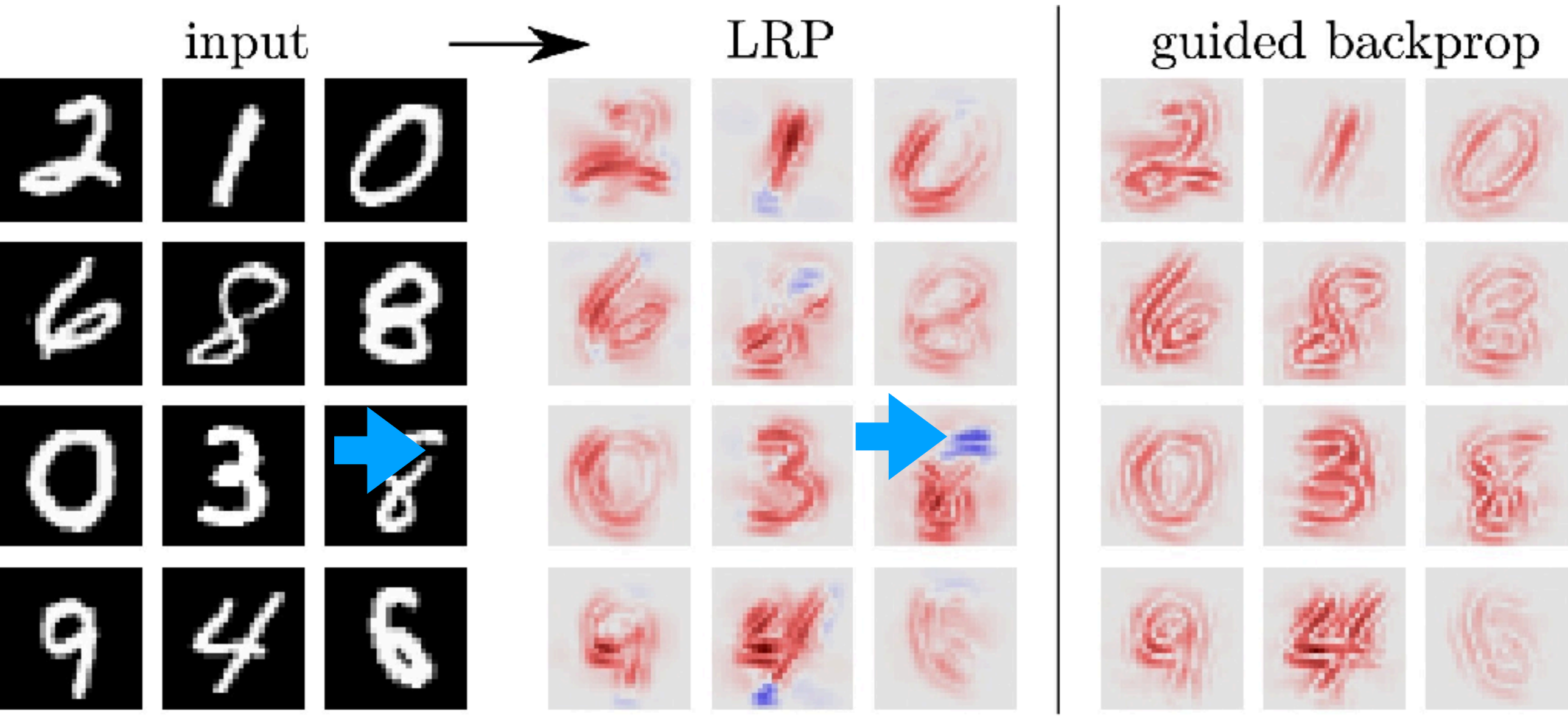
Layer-wise relevance propagation (LRP) -warstwowa propagacja istotności (Bach et al. 2015)



Layer-wise relevance propagation (LRP) -warstwowa propagacja istotności (Bach et al. 2015)



Layer-wise relevance propagation (LRP) -warstwowa propagacja istotności (Bach et al. 2015)



LRP i głębokie rozwinięcie Taylora

- można pokazać że dla sieci zbudowanych z jednostek ReLu reguła LRP-$\alpha_1\beta_0$ odpowiada iteracyjnemu obliczaniu rozwinięcia Taylora dla istotności na kolejnych warstwach.

- sama reguła redystrybucji istotności działa w praktyce także dla innych typów jednostek i architektur, ale jeszcze nie dla wszystkich wykazano, że da się je interpretować w sensie rozwinięć Taylora

Kuchnia: Struktura DNN ułatwiająca jej wyjaśnienie

- dla sieci CONV z jednostkami ReLu:

  - stosuj możliwie mało warstw FC → w przeciwnym razie istotność bardzo się „rozcieńcza” i traci związek z abstrakcyjnym konceptem, który klasyfikujemy

  - trenuj warstwy z dropoutem → to poprawia zgodność filtrów reprezentowanych przez poszczególne neurony z istotnymi cechami

  - w warstwach liniowych (CONV lub FC) ogranicz obciążenia do niedodatnich → to prowadzi do bardziej rzadkich sieci i dalej zapobiega „rozcieńczaniu” istotności

Kuchnia: wybór reguły LRP

- najpierw warto spróbować $LRP - \alpha_1\beta_0$ bo ma dobre uzasadnieni teoretyczne

- Jeśli potrzeba uwzględnić istnie cech negujących daną klasyfikację, lub mapy termiczne są zbyt „rozlane” warto spróbować $LRP - \alpha_2\beta_1$ w warstwach ukrytych

- przykładowy kod do obliczania LRP można znaleźć na http://heatmapping.org/tutorial.

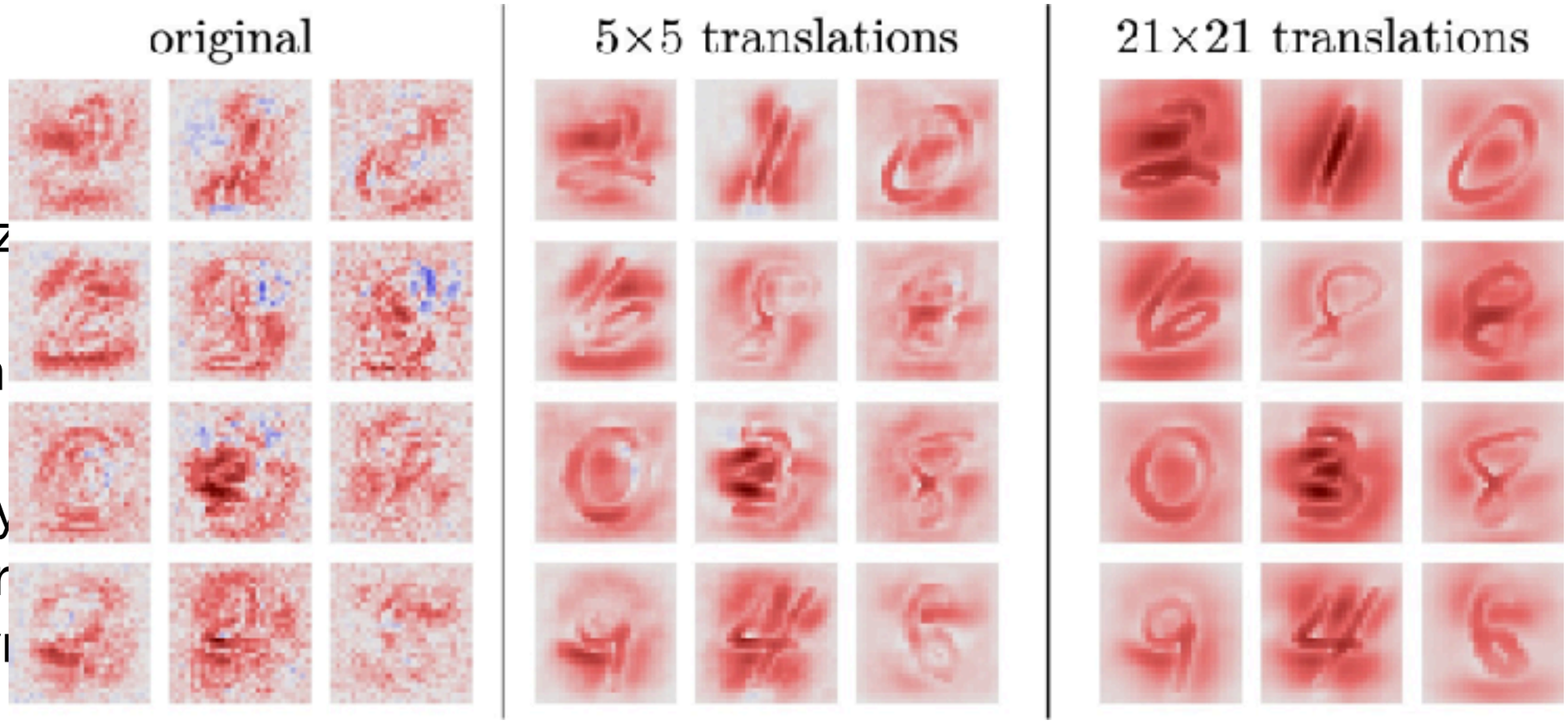
Odszumianie map termicznych

- w przypadku klasyfikatorów, które nie są optymalnie wyszkolone
- lub nie posiadają odpowiedniej struktury,
- mapy cieplne mają nieestetyczne cechy. Może to być spowodowane na przykład obecnością zaszumionych filtrów pierwszej warstwy lub dużym parametrem kroku w pierwszej warstwie splotu.

- Efekty te można złagodzić, rozważając istotność nie jednego obrazu wejściowego, ale wielu nieznacznie przesuniętych wersji obrazu. Mapy termiczne dla tych przesuniętych wersji są następnie uśredniane z wagami odwrotnie proporcjonalnymi do przesunięcia:

$$\boldsymbol{R}^{\star}(\boldsymbol{x}) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \tau^{-1}(\boldsymbol{R}(\tau(\boldsymbol{x})))$$

Odszumianie map

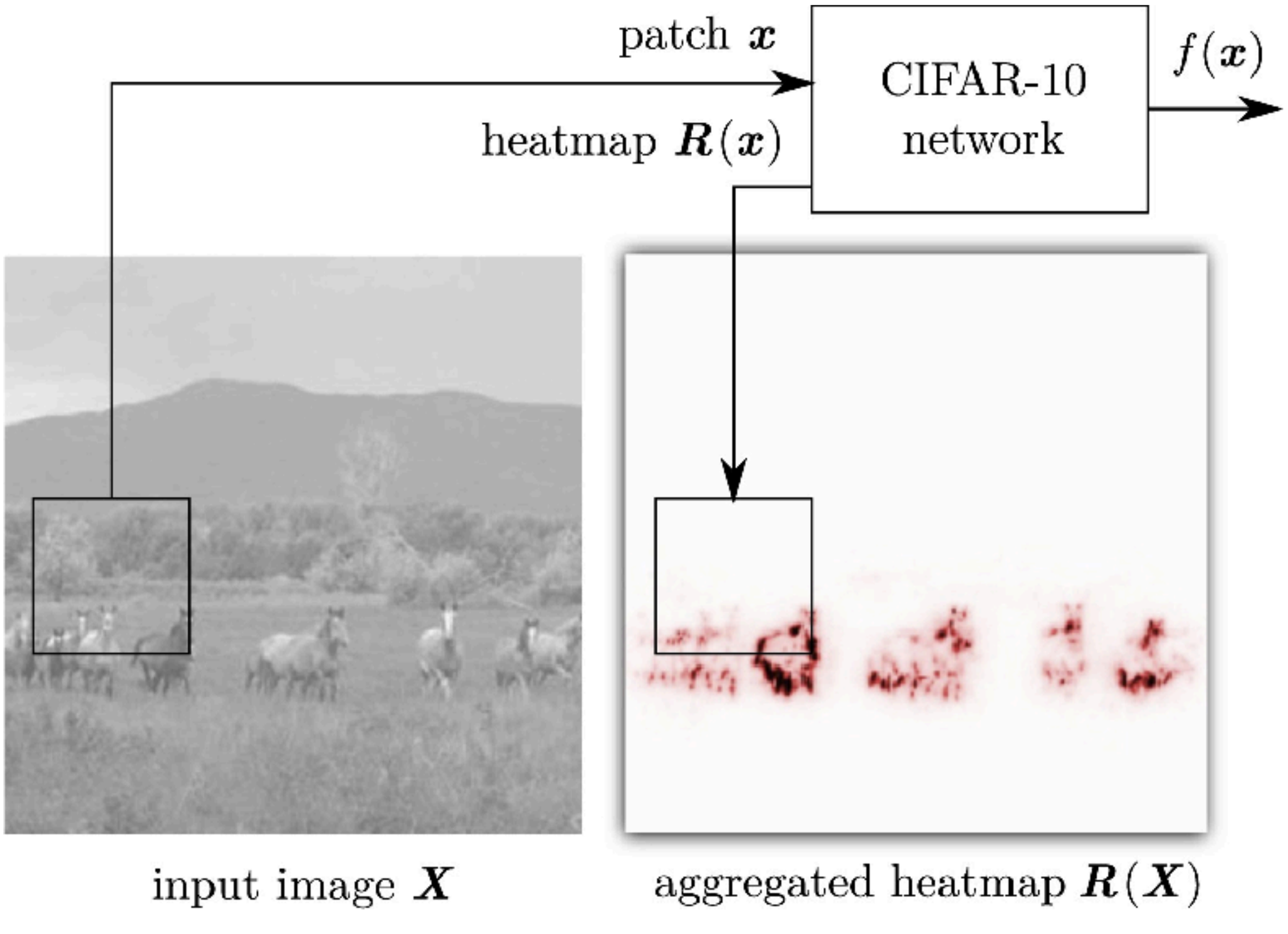


- w prz
- lub n
- mapy
na pr
duży

- Efekty te można złagodzić, rozważając istotność nie jednego obrazu wejściowego, ale wielu nieznacznie przesuniętych wersji obrazu. Mapy termiczne dla tych przesuniętych wersji są następnie uśredniane z wagami odwrotnie proporcjonalnymi do przesunięcia:

$$\boldsymbol{R}^{\star}(\boldsymbol{x}) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \tau^{-1}(\boldsymbol{R}(\tau(\boldsymbol{x})))$$

Skanowanie dużych obrazków a pomocą okienka



Przykłady zastosowań

Walidacja modelu - ocena sensu przez człowieka

target class `sci.space`.

(a)

SVM/BoW classifier

on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in

CNN/word2vec classifier

on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in

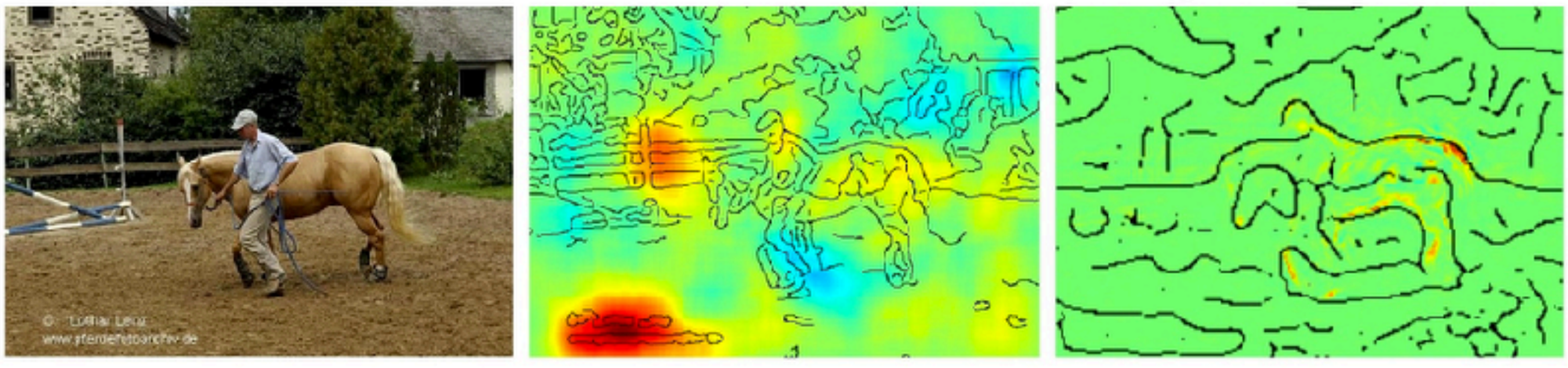
Based on Arras et al. (2016) "What is relevant in a text document? an interpretable ML approach"

(b)

input image

"horse" classification by Fisher vectors

"horse" classification by Deep neural networks



Based on Lapuschkin et al. (2016) "Analyzing classifiers: Fisher vectors and deep neural nets"

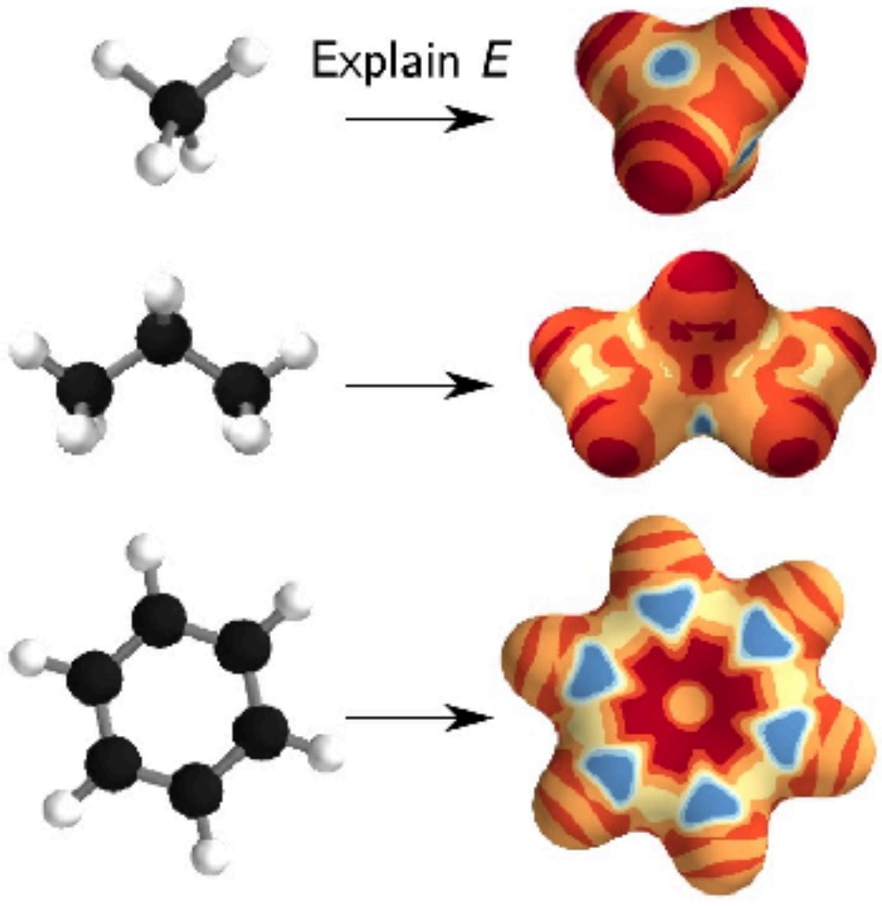
Oba klasyfikatory mają podobną skuteczność na zdjęciach koni.
Niespodziewane użycie tagu copyright do klasyfikacji i fragmentu ogrodzenia

Zastosowania do uzyskiwania wglądu w problemy naukowe: modelowanie relacji struktura-właściwości

Modele zostały wytrenowane w oparciu o dane, bez udziału symulacji fizyki w przewidywaniu.

Schütt i in. zaproponował model głębokiej sieci neuronowej, który zawiera wystarczającą strukturę i moc reprezentacyjną, aby jednocześnie osiągnąć wysoką moc predykcyjną i wyjaśnienia.

Wykorzystując analizę zaburzeń ładunku testowego (wariant analizy wrażliwości, w którym mierzy się wpływ wstawienia ładunku w danym miejsc na wyjście sieci neuronowej), stworzono trójwymiarowe mapy odpowiedzi, które podkreślają dla każdej cząsteczki struktury przestrzenne, które były najbardziej odpowiednie do wyjaśnienia modelowanej relacji struktura-właściwość.



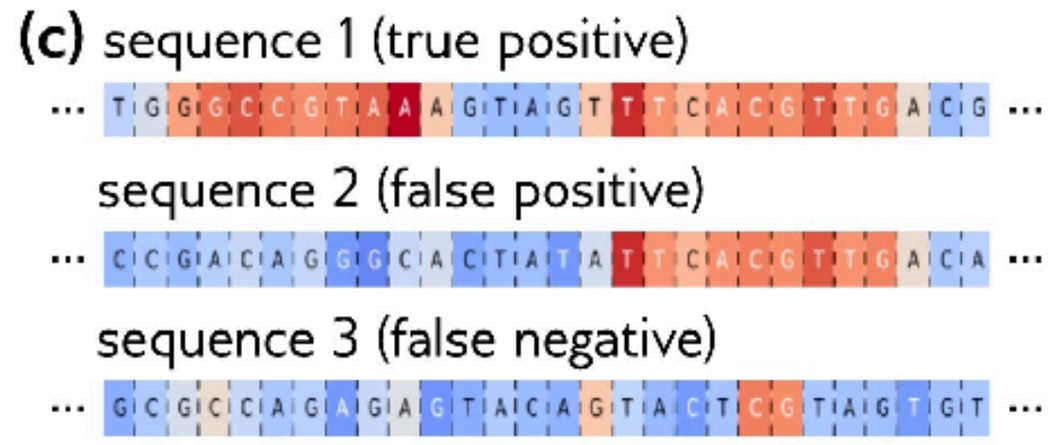
Based on Schütt et al. (2017) "Quantum-chemical insights from deep tensor neural networks"

Zastosowania do uzyskiwania wglądu w problemy naukowe: mapowania sekwencji DNA na miejsca wiązania

Alipanahi i in. wytrenował sieć konwolucyjną do mapowania sekwencji DNA na miejsca wiązania białka.

Za jej pomocą testowano jakie nukleotydy z tej sekwencji są najbardziej istotne dla wyjaśnienia obecności tych miejsc wiązania.

Wykorzystali analizę opartą na zaburzeniach, w której mierzy się istotność każdego nukleotydu na podstawie wpływu mutacji na prognozę sieci neuronowej.



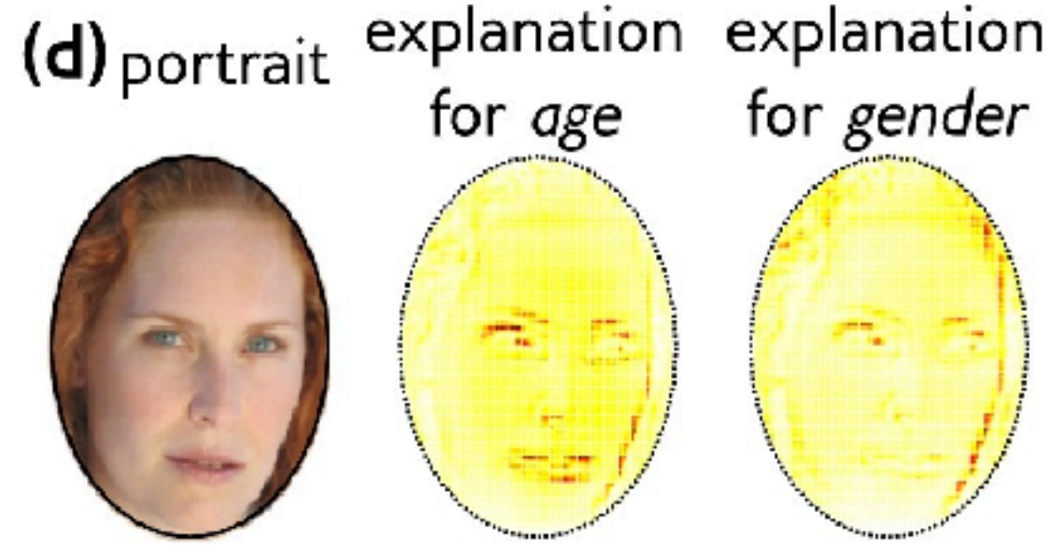
Adapted from Vidovic et al. (2016) "Feature importance measure for non-linear learning algorithms"

Zastosowania do uzyskiwania wglądu w problemy naukowe

- Techniki wyjaśniania mają również potencjalne zastosowanie w analizie obrazów twarzy.

- Ich bezpośrednia interpretacja pod kątem rzeczywistych cech obrazu wejściowego może być trudna.

- Arbabzadah i in. zastosował technikę LRP, aby zidentyfikować, które piksele na danym obrazie są odpowiedzialne za wyjaśnienie, na przykład, atrybutów wieku i płci.



Based on Arbabzadah et al. (2016) "Identifying individual facial expressions by deconstructing a neural network"

więcej o tej technice:

http://www.heatmapping.org/

toolbox w Keras z implementacją tej oraz innych technik

https://github.com/albermax/innvestigate