

JAK SIECI NEURONOWE WIDZĄ ŚWIAT

OD PIKSELI DO ATAKÓW ADWERSARIALNYCH

Marzena Halama

INSTYTUT INFORMATYKI TEORETYCZNEJ I STOSOWANEJ
POLSKIEJ AKADEMII NAUK

Agenda

- 01** Poznajmy się
- 02** Czym jest sztuczna inteligencja
- 03** Zobaczyć świat z perspektywy AI
- 04** Wyzwania, możliwości i zagrożenia



INSTYTUT INFORMATYKI TEORETYCZNEJ
I STOSOWANEJ POLSKIEJ AKADEMII NAUK



Obecnie pracuje jako Asystent Profesora - piszę doktorat
o zastosowaniu LLM-ów w Chemii

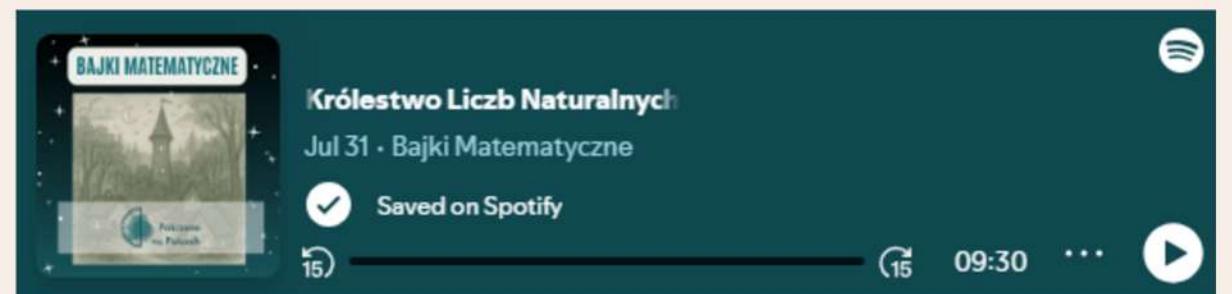
Skończyłam Matematykę Stosowaną na Politechnice Śląskiej w Gliwicach

Pracowałam jako Software Quality Assurance Tester w 3 SOFT S.A. Katowice

Popularyzuje naukę w projekcie: „Policzone na Palcach”

Nagrywam Bajki Matematyczne, które łączą świat liczb z wyobraźnią poprzez opowieści
i emocje - podcast dostępny na spotify

SCAN ME



**mgr Marzena
Halama**



Marzena Halama

KTÓRA STRUKTURA ODPOWIADA ZA INTERPRETACJĘ BODŹCÓW WZROKOWYCH?

Nerw wzrokowy

A

Soczewka

Siatkówka

C

D

Kora wzrokowa
w płacie potylicznym

KTÓRA STRUKTURA ODPOWIADA ZA INTERPRETACJĘ BODŹCÓW WZROKOWYCH?

Nerw wzrokowy

A

Soczewka

Siatkówka

C

Kora wzrokowa
w płacie potylicznym

**OCZY NIE „WIDZĄ” W SENSIE INTERPRETACJI OBRAZU
ONE TYLKO REJESTRUJĄ ŚWIATŁO I PRZEKAZUJĄ DANE DALEJ**



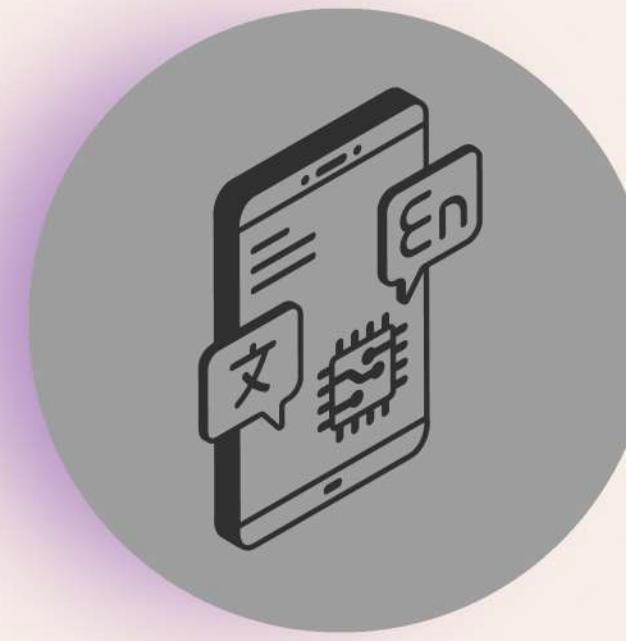
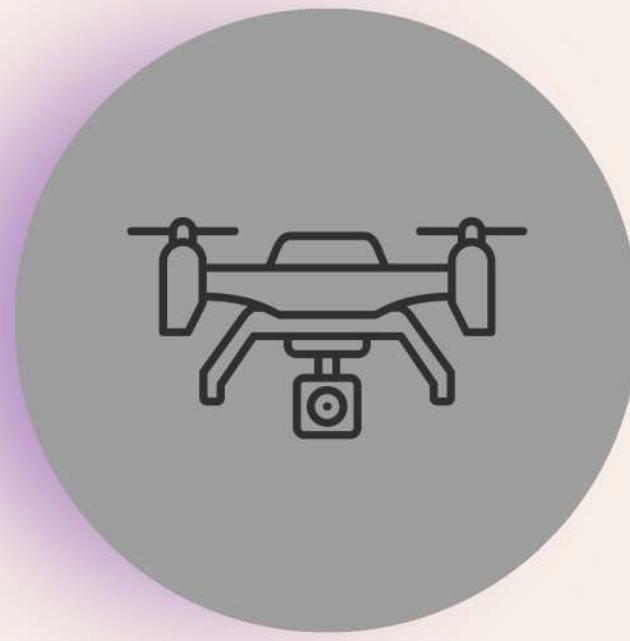
**PRAWDZIWE „WIDZENIE” ZACHODZI DOPIERO W KORZE WZROKOWEJ,
ZNAJDUJĄCEJ SIĘ W PŁACIE POTYLICZNYM MÓZGU...**

A SZTUCZNE SIECI NEURONOWE PRÓBUJĄ ROBIĆ TO SAMO TYLKO CYFROWO.

Marzena Halama

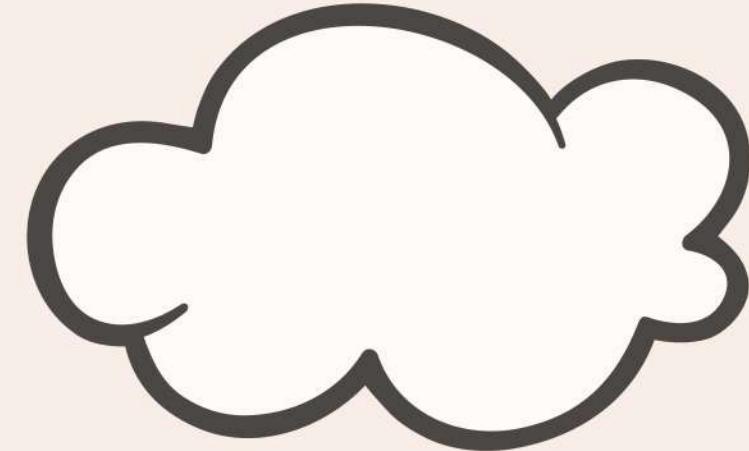
CO TO JEST SZTUCZNA INTELIGENCJA

Sztuczna inteligencja to dziedzina wiedzy i technologii, której celem jest tworzenie systemów zdolnych do wykonywania zadań, które normalnie wymagają ludzkich zdolności poznawczych, takich jak rozumienie języka, uczenie się, rozwiązywanie problemów, percepcja wzrokowa czy podejmowanie decyzji [1].



Marzena Halama

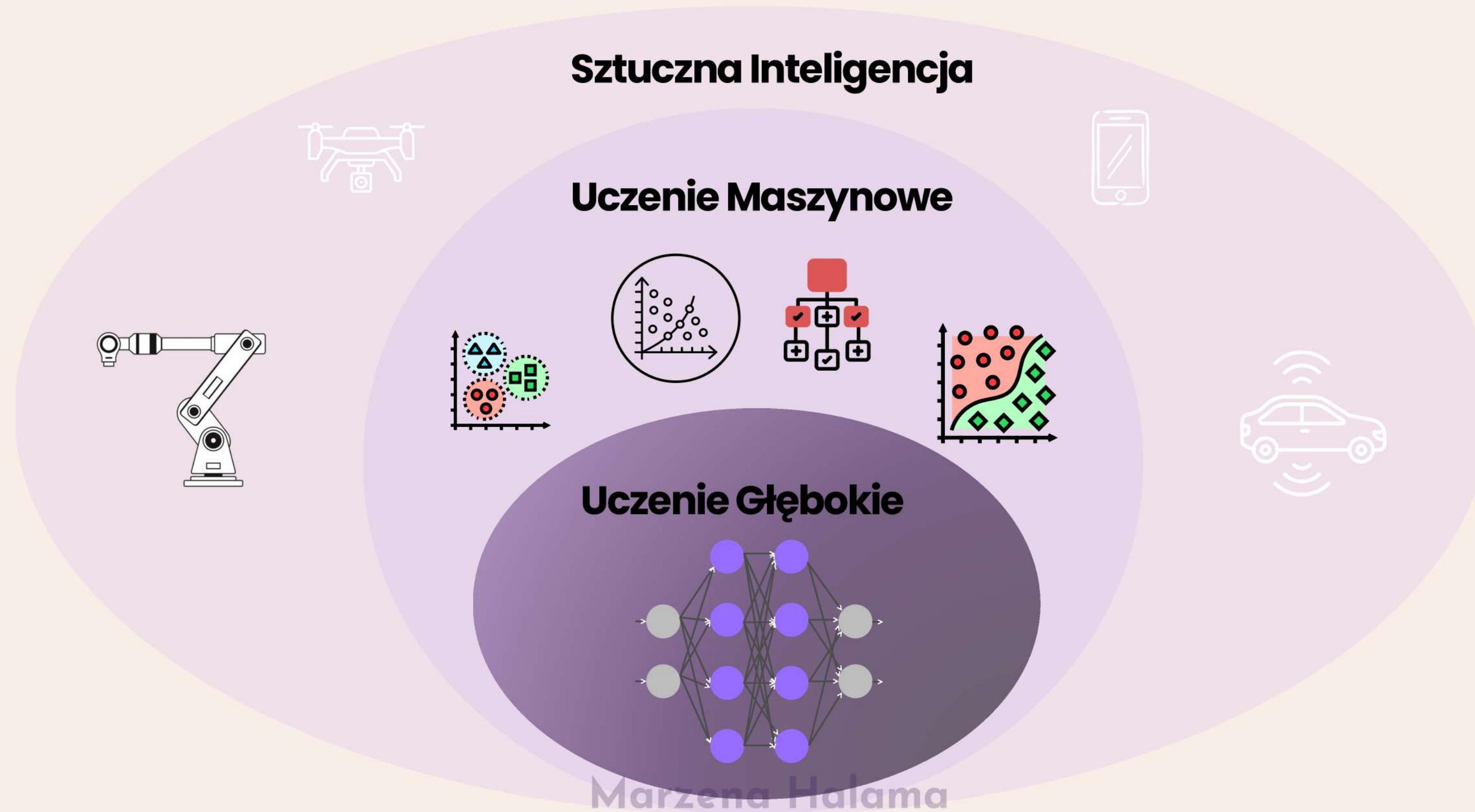
[1] Russell, Stuart J., and Peter Norvig. Artificial intelligence: a modern approach. Pearson, 2016.



DOKĄD ZMIERZA ROZWÓJ TECHNOLOGICZNY

Marzena Halama

DIAGRAM PRZEDSTAWIAJĄCY HIERARCHICZNĄ ZALEŻNOŚĆ POMIĘDZY SZTUCZNĄ INTELIGENCJĄ, UCZENIEM MASZYNOWYM
I GŁĘBOKIMI SIECIAMI NEURONOWYMI.



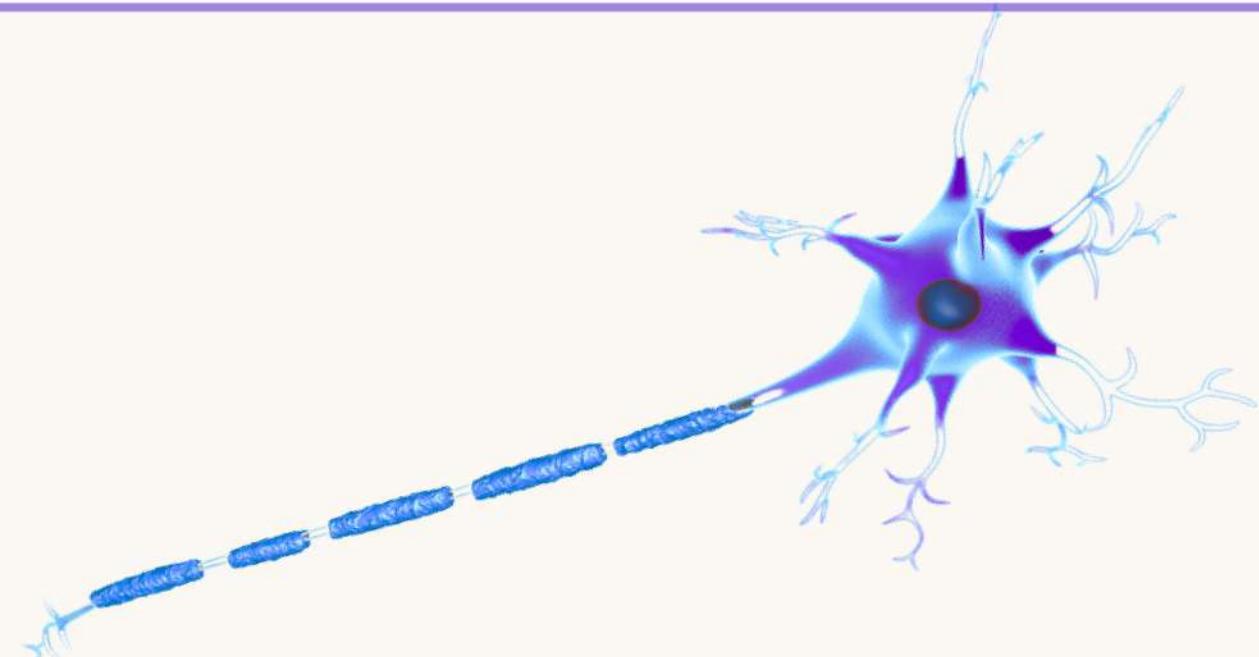
To sposób, w jaki komputer uczy się rozpoznawać wzorce

Marzena Halama

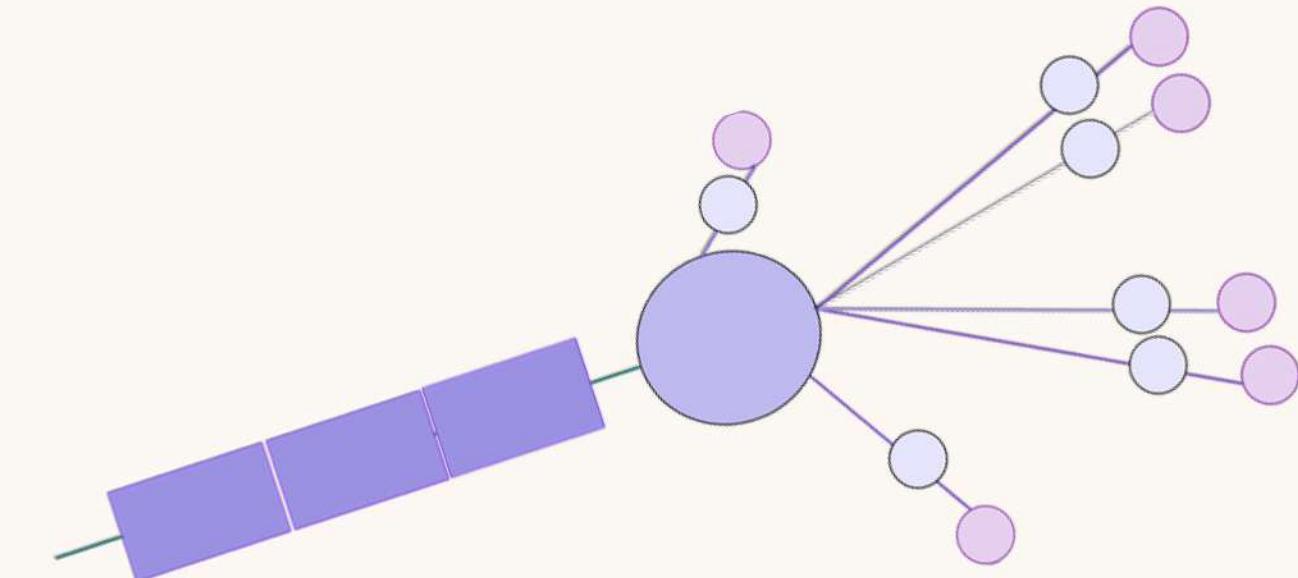
UCZENIE GŁĘBOKIE

Dzieje się tak dzięki sztucznym sieciom neuronowym, które naśladują sposób działania neuronów w ludzkim mózgu

neuron w mózgu człowiek



sztuczny neuron (perceptron)

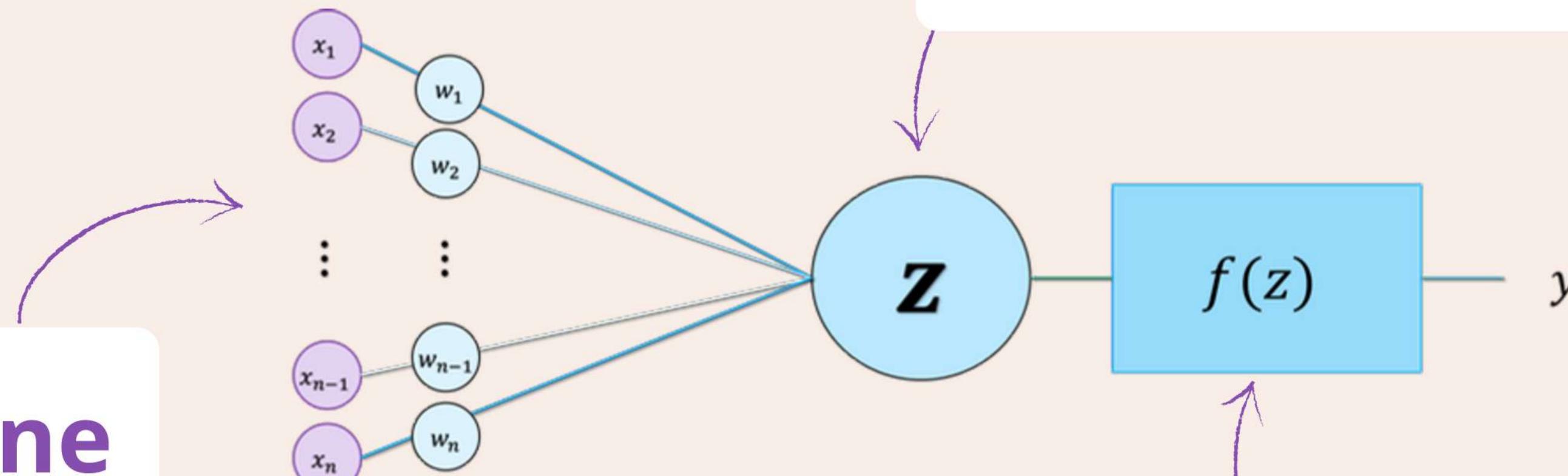


PERCEPTRON

dane

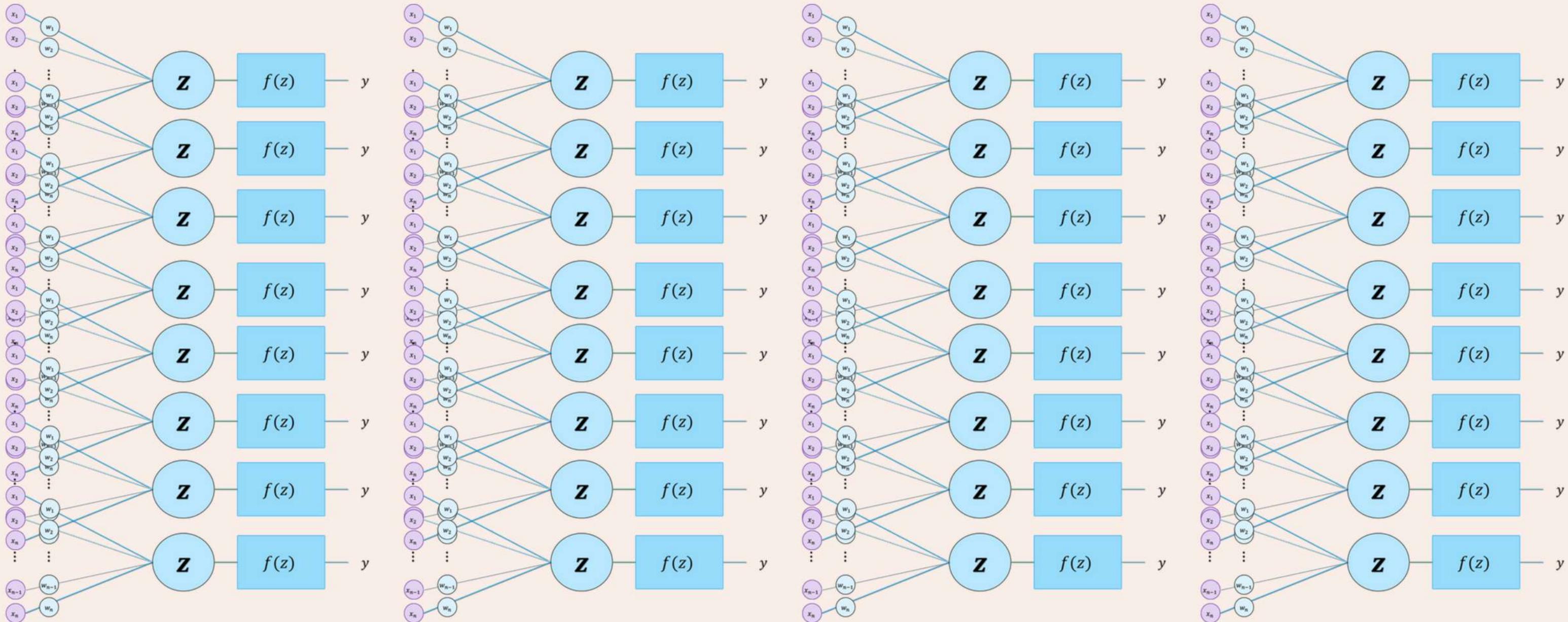


Przykładowe zadanie dla perceptronów:
Dane wejściowe: (x₁, x₂, ..., x_n)
Wagi: (w₁, w₂, ..., w_n)
Wynik: f(x) = y



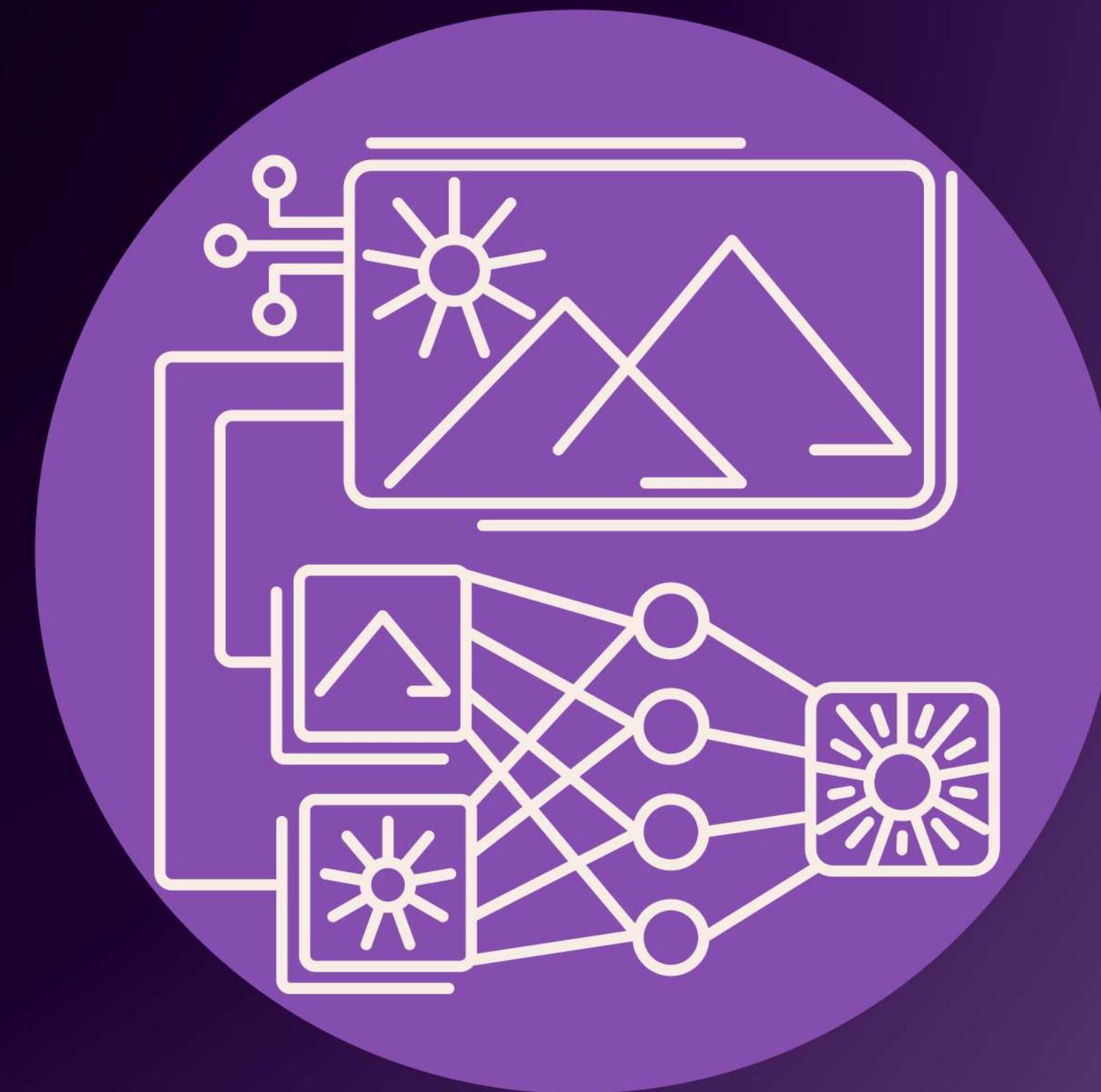
Marzena Halama

SZTUCZNA SIEĆ NEURONOWA



Marzena Halama

ANALIZA OBRAZÓW



Marzena Halama

ZADANIA ZWIĄZANE Z PRZETWARZANIA OBRAZÓW



KLASYFIKACJA OBIEKTÓW
(ANG. OBJECT CLASSIFICATION)



LOKALIZACJA OBIEKTÓW
(ANG. OBJECT LOCATION)



WYKRYWANIE OBIEKTÓW
(ANG. OBJECT DETECTION)

Marzena Halama

ANALIZA STRUKTURY OBRAZU

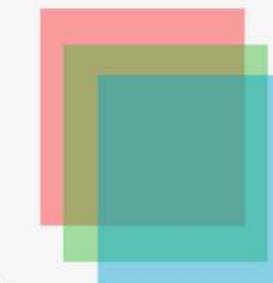


Każdy obraz cyfrowy składa się z pikseli

ANALIZA STRUKTURY OBRAZU



Każdy obraz cyfrowy składa się z pikseli



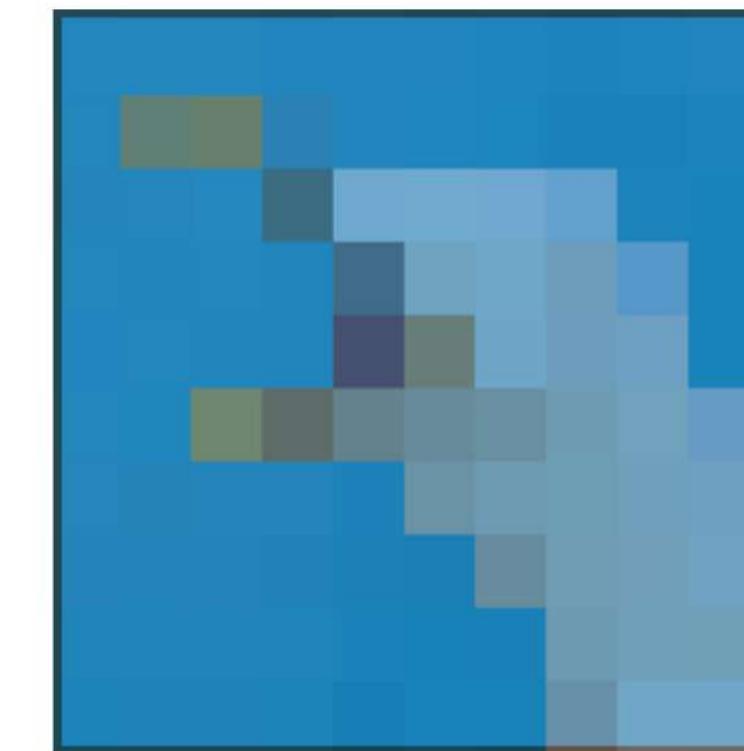
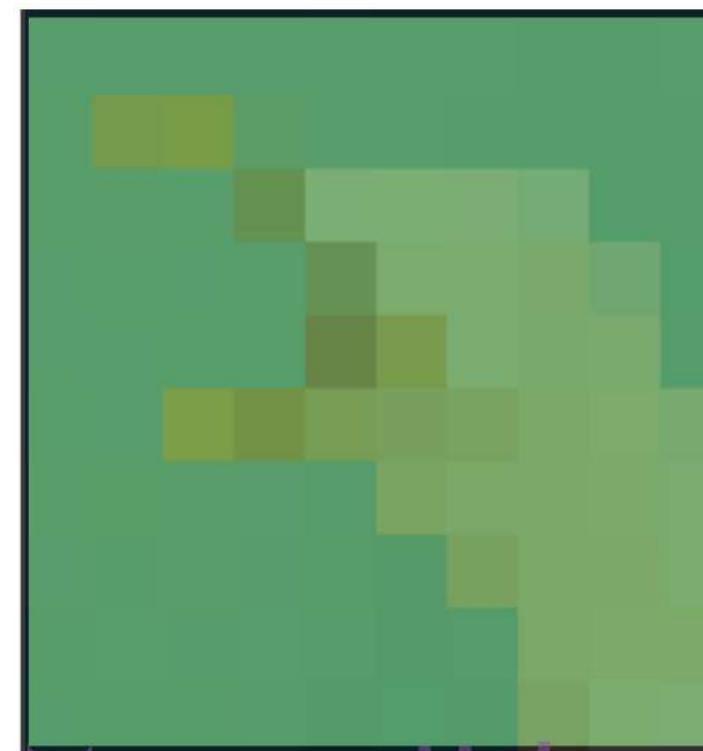
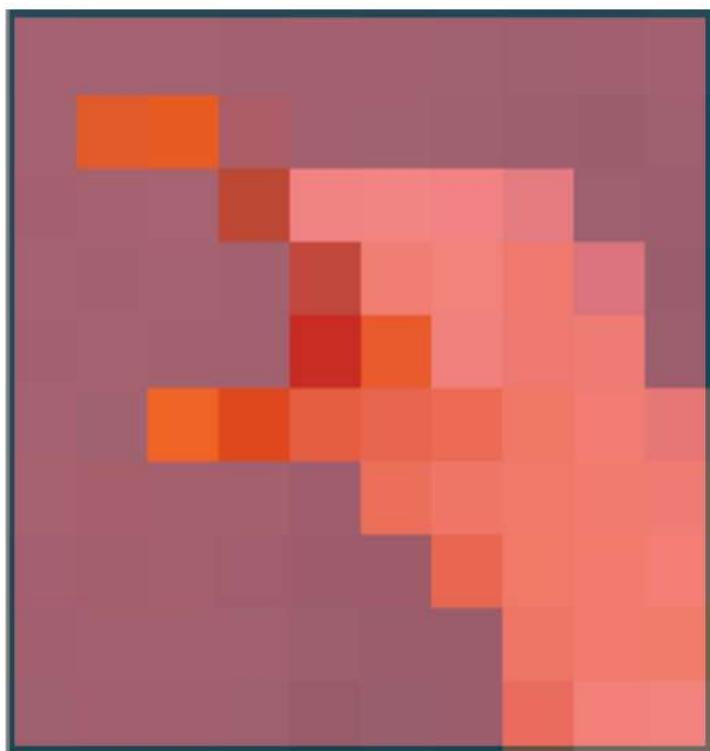
Każdy piksel składa się z trzech kanałów



Każdy kolor w modelu RGB opisujemy jako trójkę liczb:

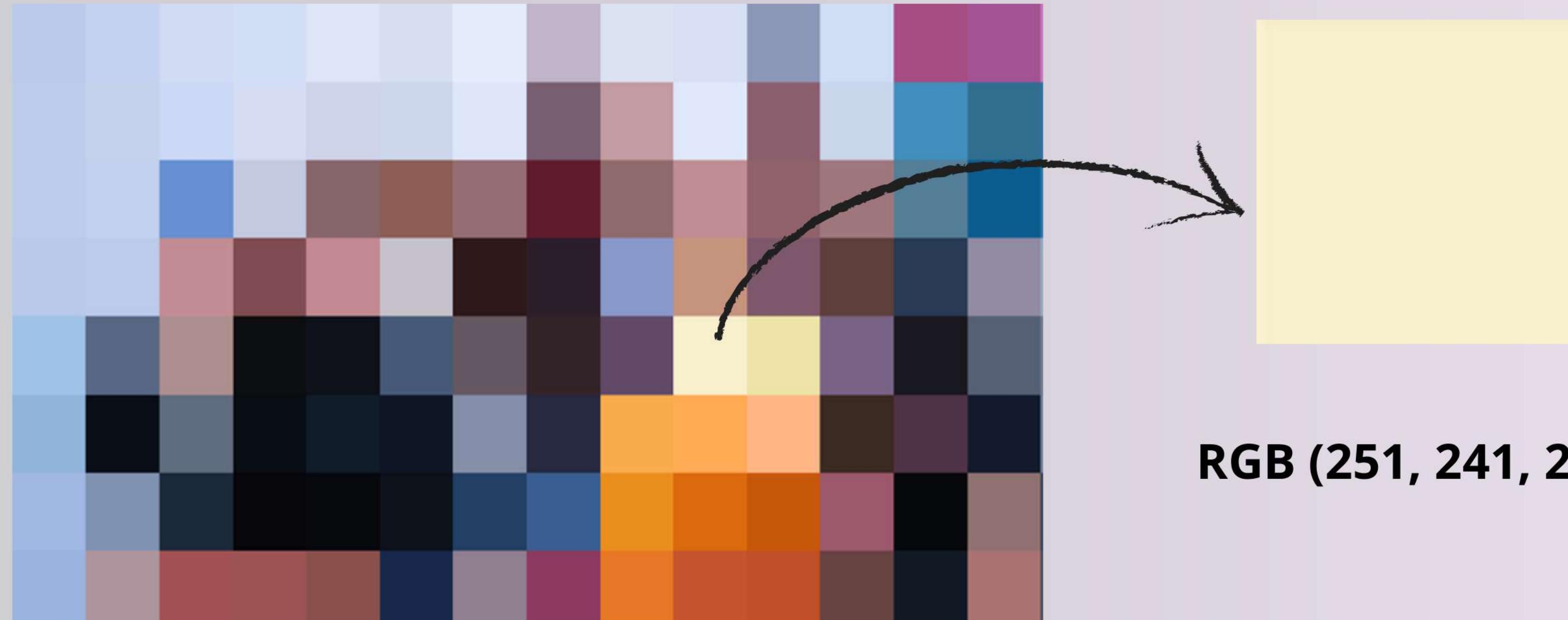
$$(R, G, B)$$

Każda liczba określa intensywność światła danego koloru od 0 do 255



Marzena Halama

KAŻDY PIKSEL OKREŚLAJĄCA INTENSYWNOŚĆ CZERWIENI, ZIELENI I BŁĘKITU



Marzena Halama

**ABY UZYSKAĆ OBRAZ CZARNO-BIAŁY
ŁĄCZYMY TE KANAŁY W JEDEN**



Marzena Halama

KAŻDY PIKSE BĘDZIE MIEĆ WARTOŚĆ OD 0 DO 255



Marzena Halama

DZIĘKI TEMU OTRZYMUJEMY OBRAZ W ODCIENIACH SZAROŚCI

57	117	97	
146	0	97	99
97	99	117	0
0	220	146	175
0	235	166	178

Marzena Halama

PIKSEL = WARTOŚĆ LICZBOWA

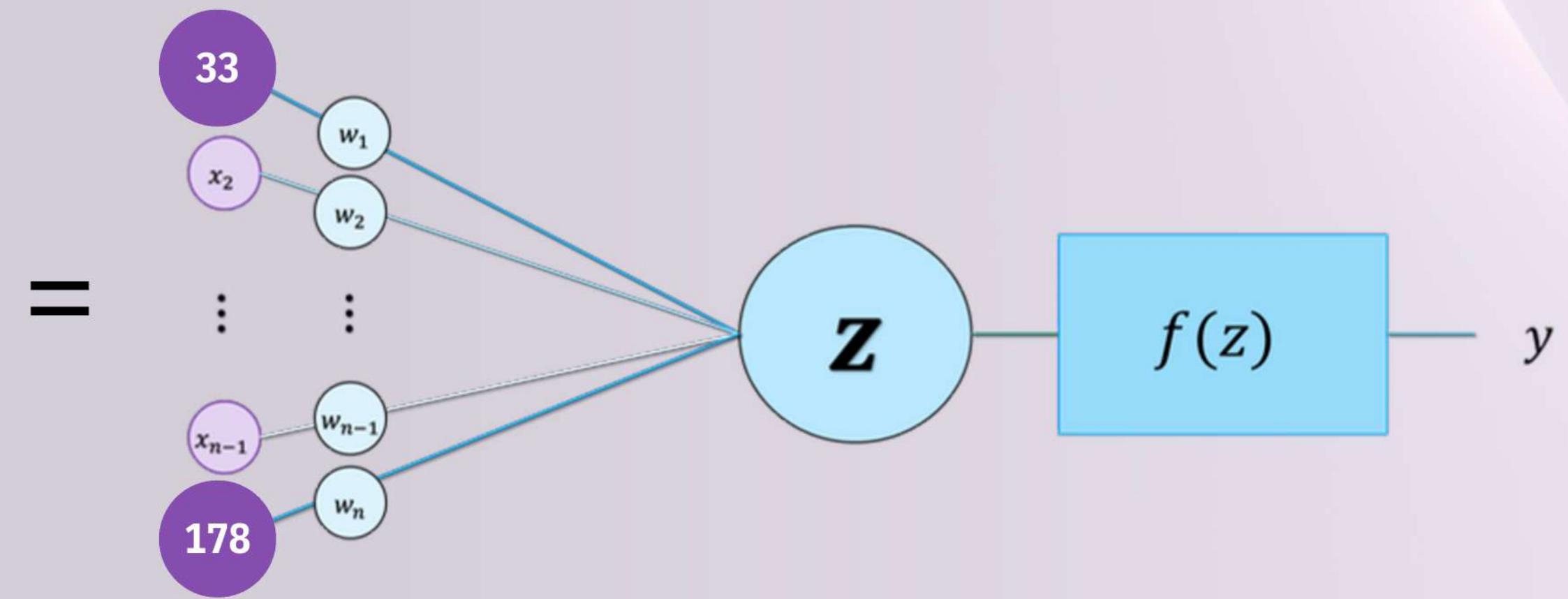
	57	117	97
146	0	97	99
97	99	117	0
0	220	146	175
0	235	166	178

=

$$A = \begin{bmatrix} 33 & 57 & 117 & 97 \\ 146 & 0 & 97 & 99 \\ 97 & 99 & 117 & 0 \\ 0 & 220 & 146 & 175 \\ 0 & 235 & 166 & 178 \end{bmatrix}$$

WARTOŚĆ LICZBOWA = WEJŚCIE PERCEPTRONU

$$A = \begin{bmatrix} 33 & 57 & 117 & 97 \\ 146 & 0 & 97 & 99 \\ 97 & 99 & 117 & 0 \\ 0 & 220 & 146 & 175 \\ 0 & 235 & 166 & 178 \end{bmatrix}$$



KONWOLUCYJNE SIECI NEURONOWE

jak komputer widzi obraz

Wartości pikseli są matematycznie przekształcane przez filtry, aby wydobyć z obrazu kształty, krawędzie i wzorce.

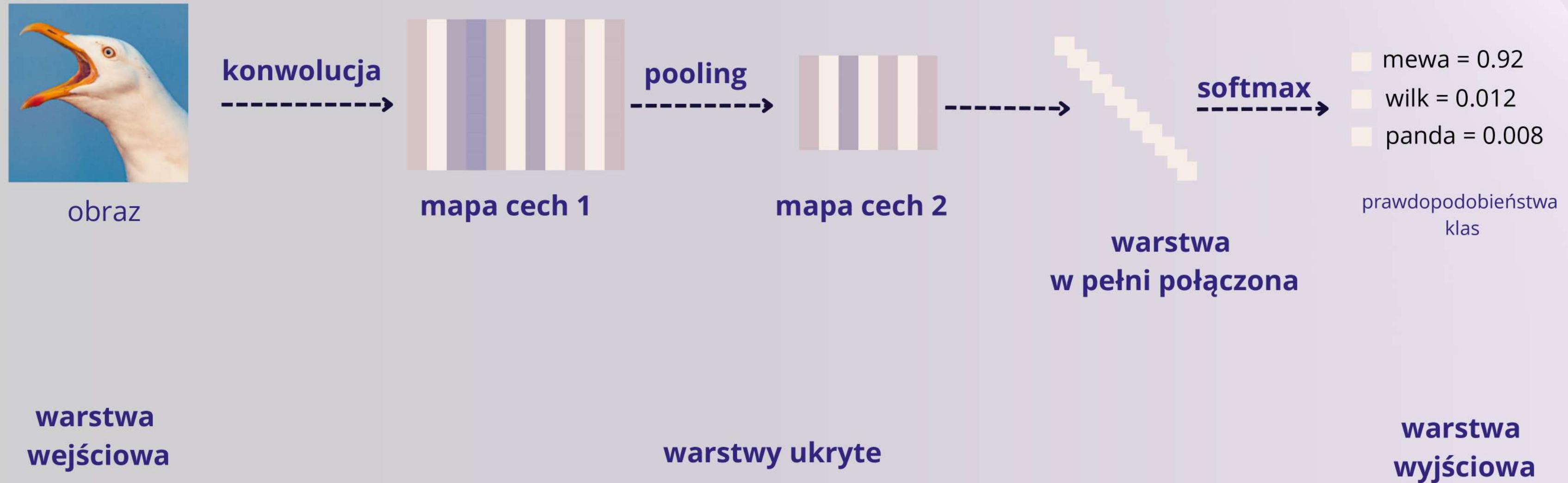
$$S(i, j) = \sum_m \sum_n I(i + m, j + n) \cdot K(m, n)$$

W ten sposób powstają mapy cech (feature maps), które reprezentują różne poziomy informacji wizualnej.

Marzena Halama

SIECI KONWOLUCYJNE (ANG. CONVOLUTIONAL NEURAL NETWORKS, CNN) ZWANE RÓWNIEŻ SIECIAMI SPŁOTOWYMI POWSTAŁY JAKO NARZĘDZIE ANALIZY I ROZPOZNAWANIA OBRAZÓW WZOROWANE NA SPOSOBIE DZIAŁANIA NASZYCH ZMYSŁÓW. BUDOWA TYCH SIECI JEST NASTĘPUJĄCA:

ARCHITEKTURA KONWOLUCYJNEJ SIECI NEURONOWEJ

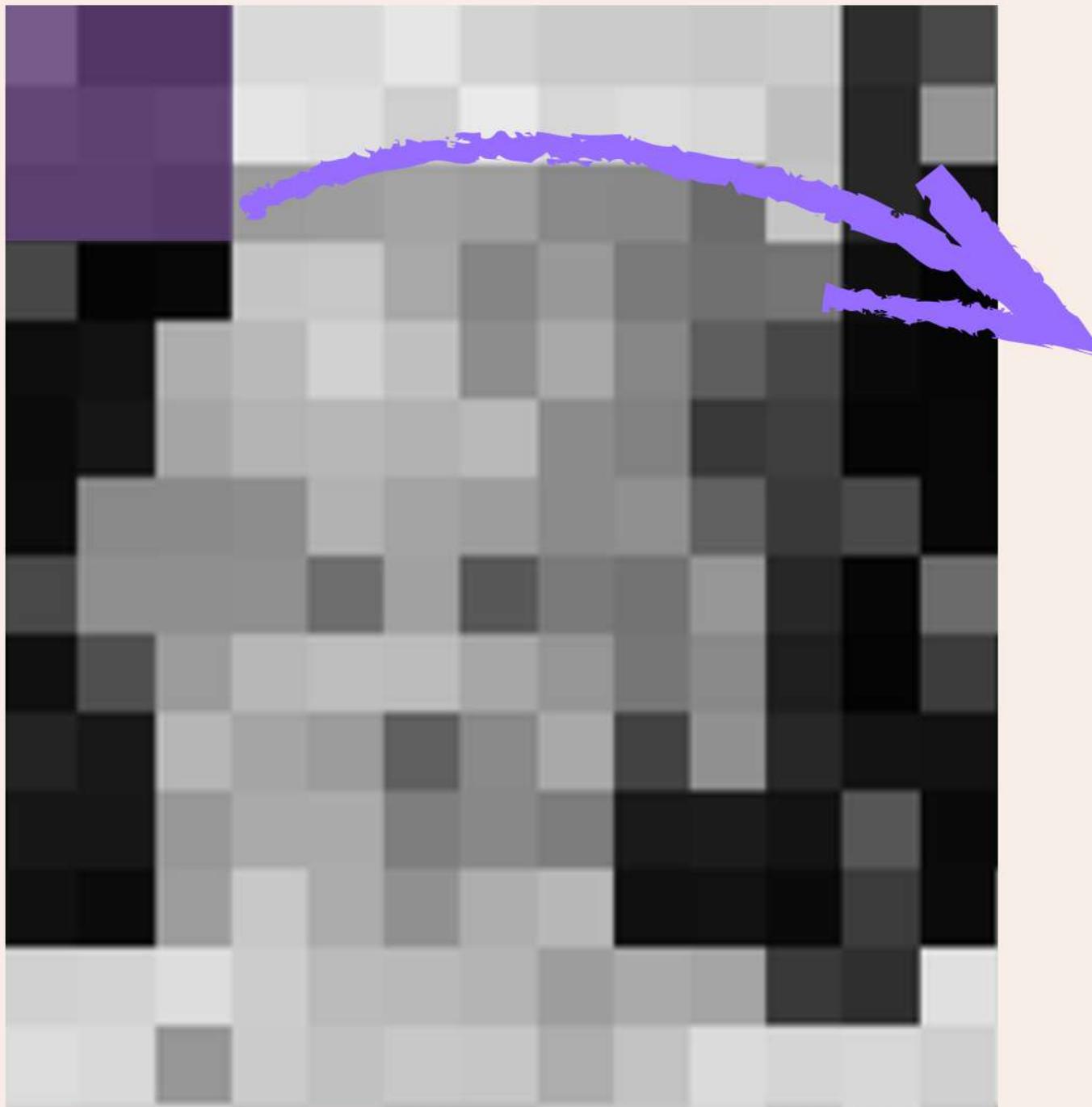


KONWOLUCYJNE SIECI NEURONOWE



Marzena Halama

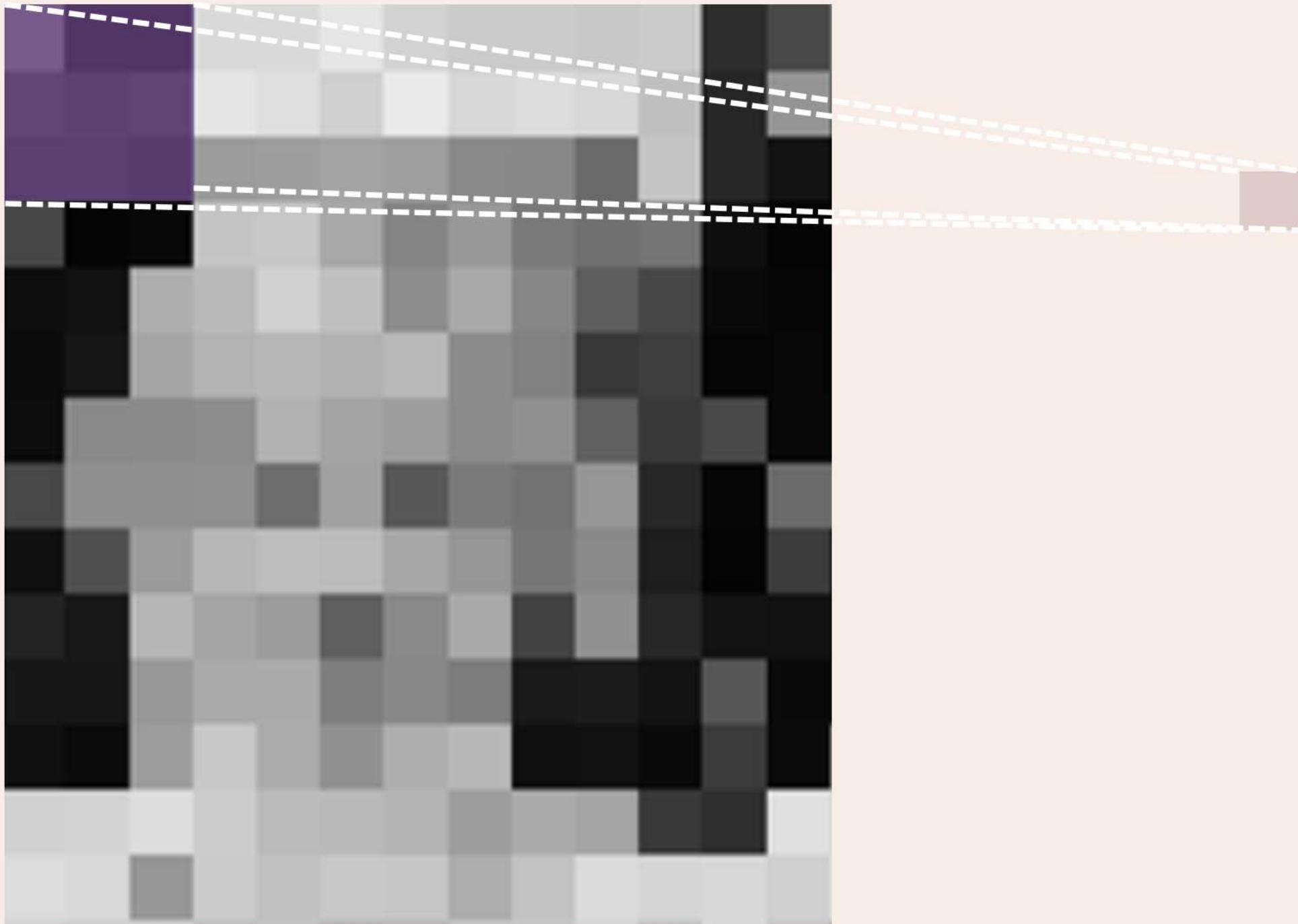
KONWOLUCYJNE SIECI NEURONOWE



Filtr przesuwa się po obrazie i tworzy
mapę cech

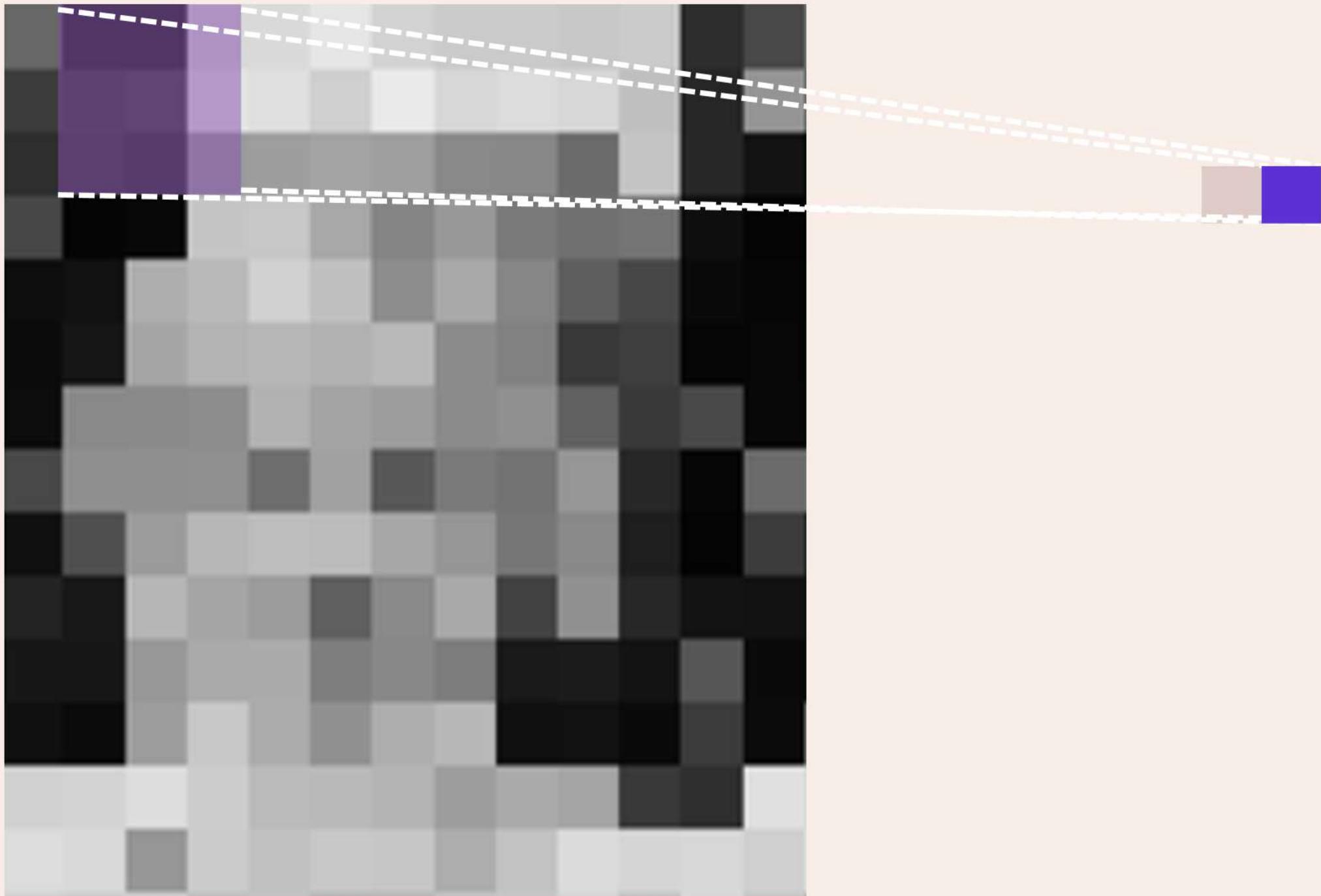
Marzena Halama

KONWOLUCYJNE SIECI NEURONOWE



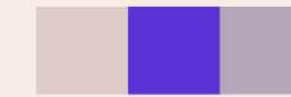
Marzena Halama

KONWOLUCYJNE SIECI NEURONOWE



Marzena Halama

KONWOLUCYJNE SIECI NEURONOWE



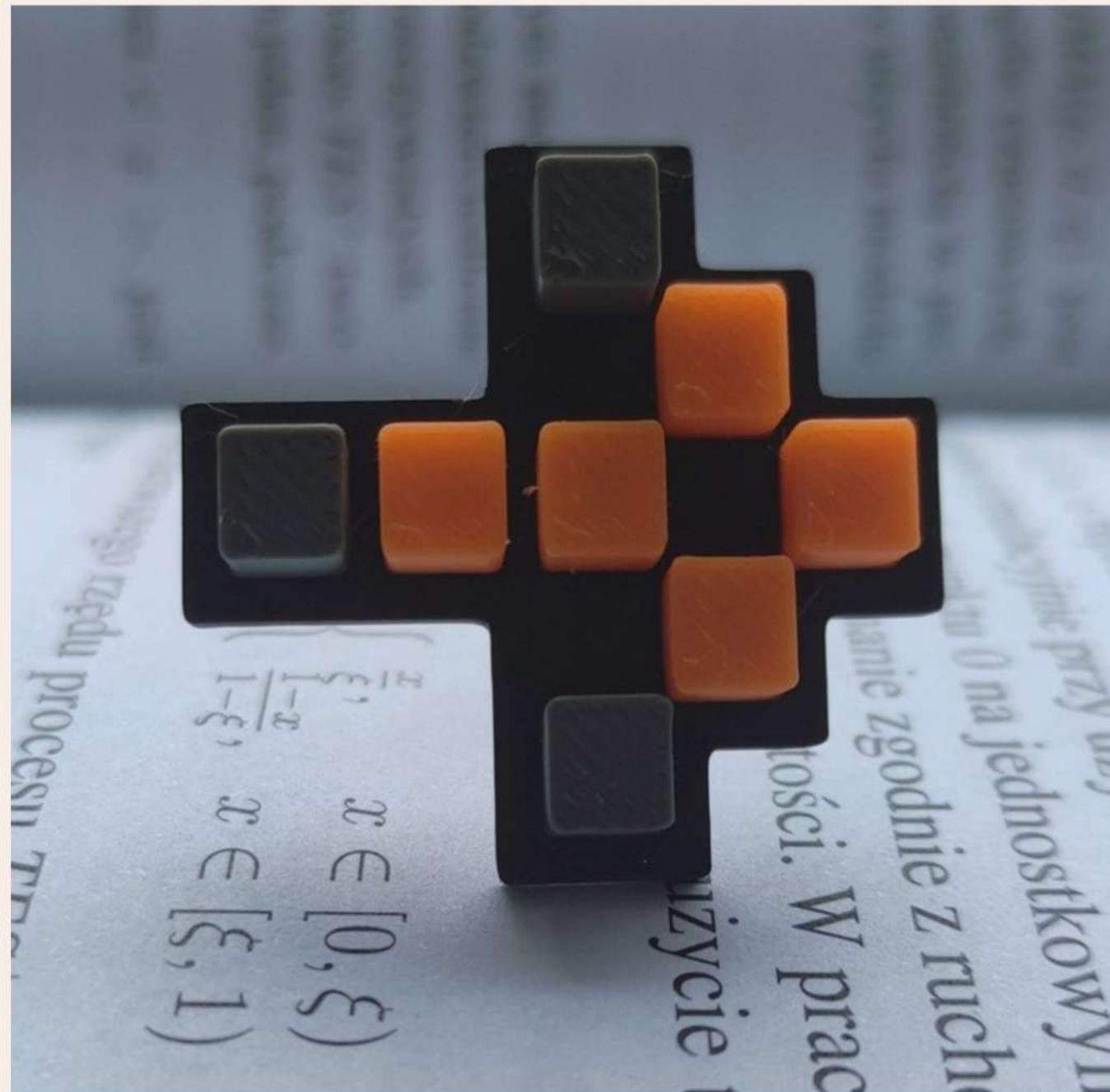
Marzena Halama

PONIŻEJ PODANO WZORY MACIERZY (KERNELI, FILTRÓW), KTÓRE ODPOWIADAJĄ POSZCZEGÓLNYM PRZEKSZTAŁCENIOM



Marzena Halama

WIZUALIZACJA DZIAŁANIA SIECI CNN - WARSTWA PO WARSTWIE - PRZYKŁD DZIAŁANIA

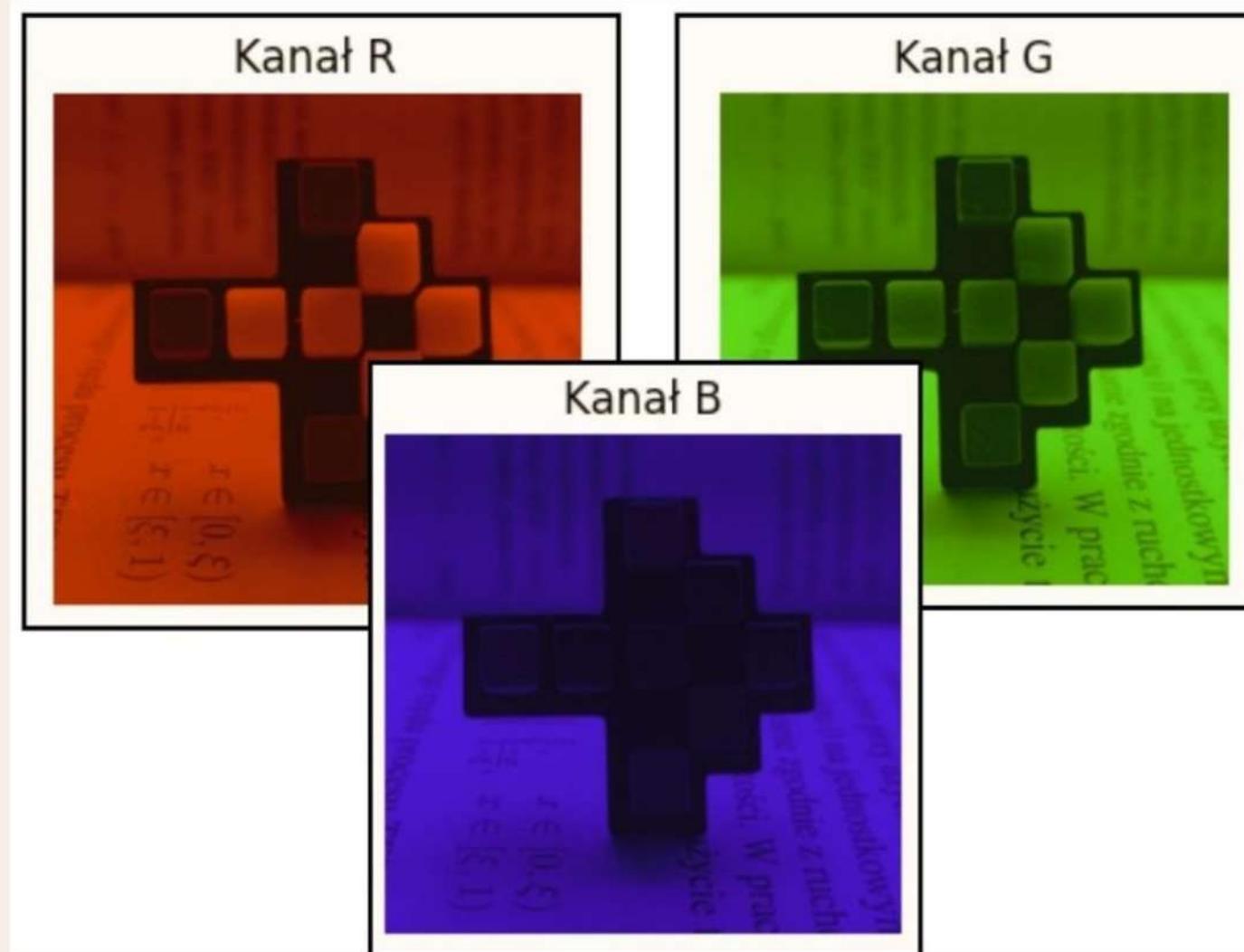


Warstwa wejściowa czyli obraz (o rozmiarze 224x224 pikseli) w sieciach konwolucyjnych jest reprezentowany jako dwuwymiarowa macierz, gdzie każdy element odpowiada wartości intensywności pikseli.

W przypadku obrazów kolorowych są to trójwymiarowe tensory, w których trzeci wymiar reprezentuje kanały kolorów.

Marzena Halama

WIZUALIZACJA DZIAŁANIA SIECI CNN - WARSTWA PO WARSTWIE - PRZYKŁD DZIAŁANIA



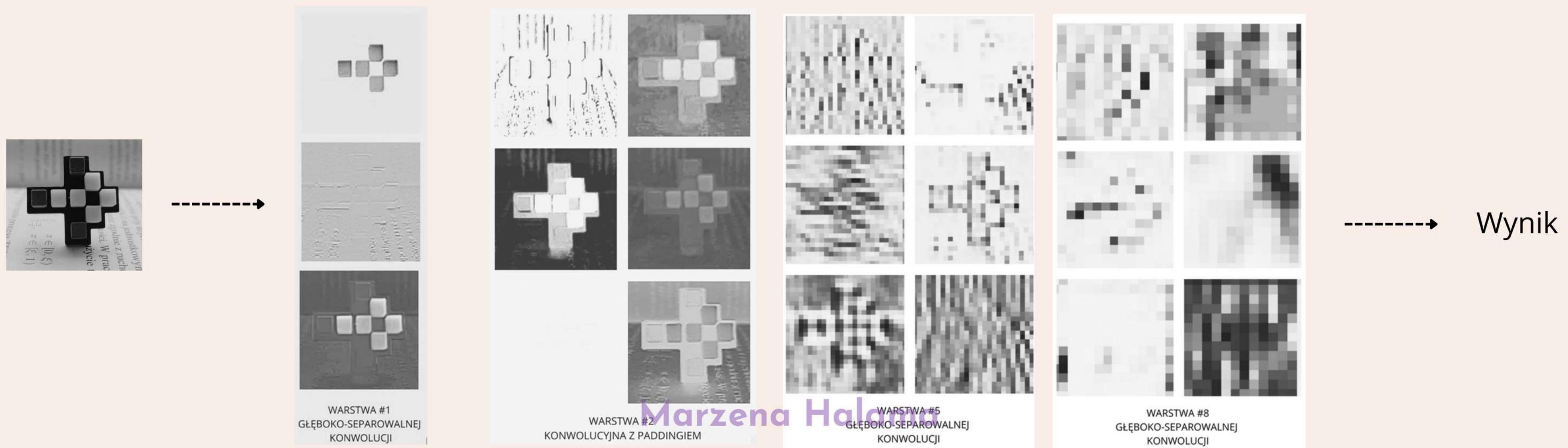
Warstwa wejściowa czyli obraz (o rozmiarze 224x224 pikseli) w sieciach konwolucyjnych jest reprezentowany jako dwuwymiarowa macierz, gdzie każdy element odpowiada wartości intensywności pikseli.

W przypadku obrazów kolorowych są to trójwymiarowe tensorzy, w których trzeci wymiar reprezentuje kanały kolorów.

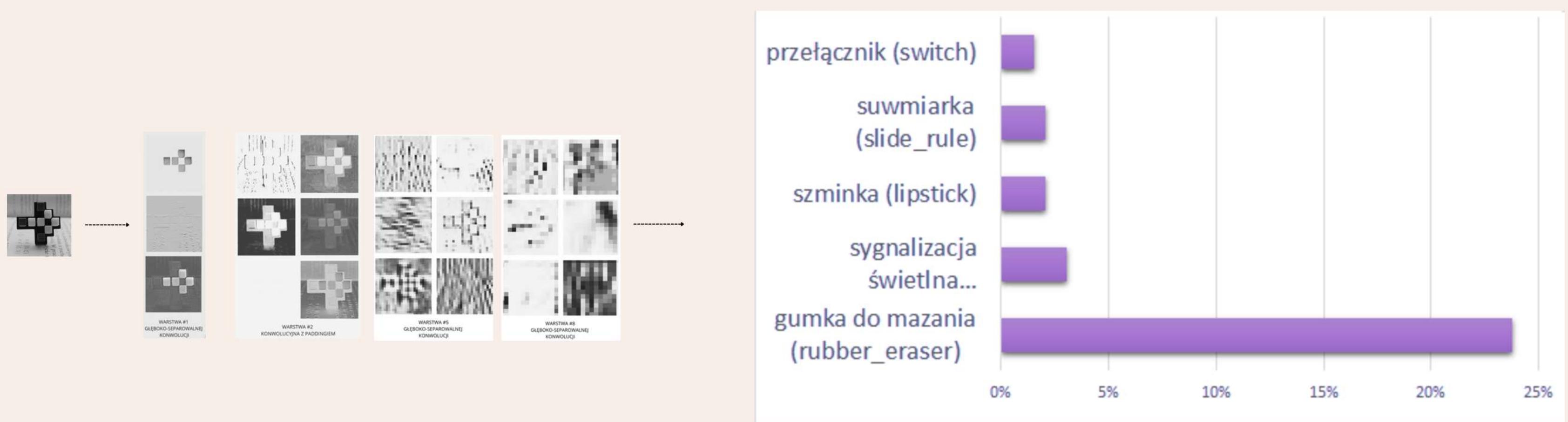
Marzena Halama

WIZUALIZACJA DZIAŁANIA SIECI CNN - WARSTWA PO WARSTWIE - PRZYKŁD DZIAŁANIA

Mapowanie cech od ogólnych do bardziej abstrakcyjnych

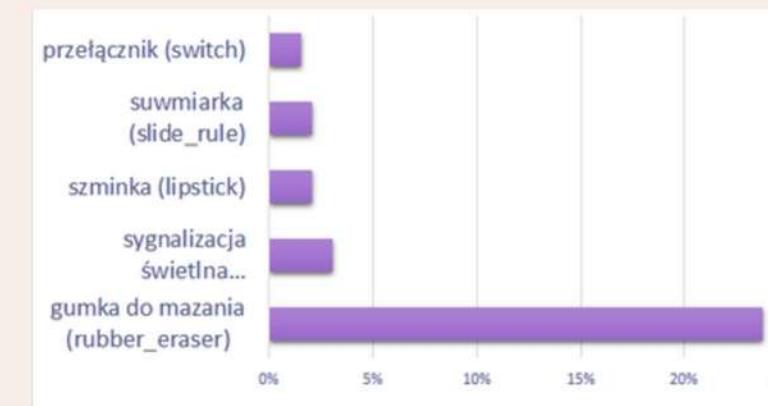


WIZUALIZACJA DZIAŁANIA SIECI CNN - WARSTWA PO WARSTWIE - PRZYKŁD DZIAŁANIA

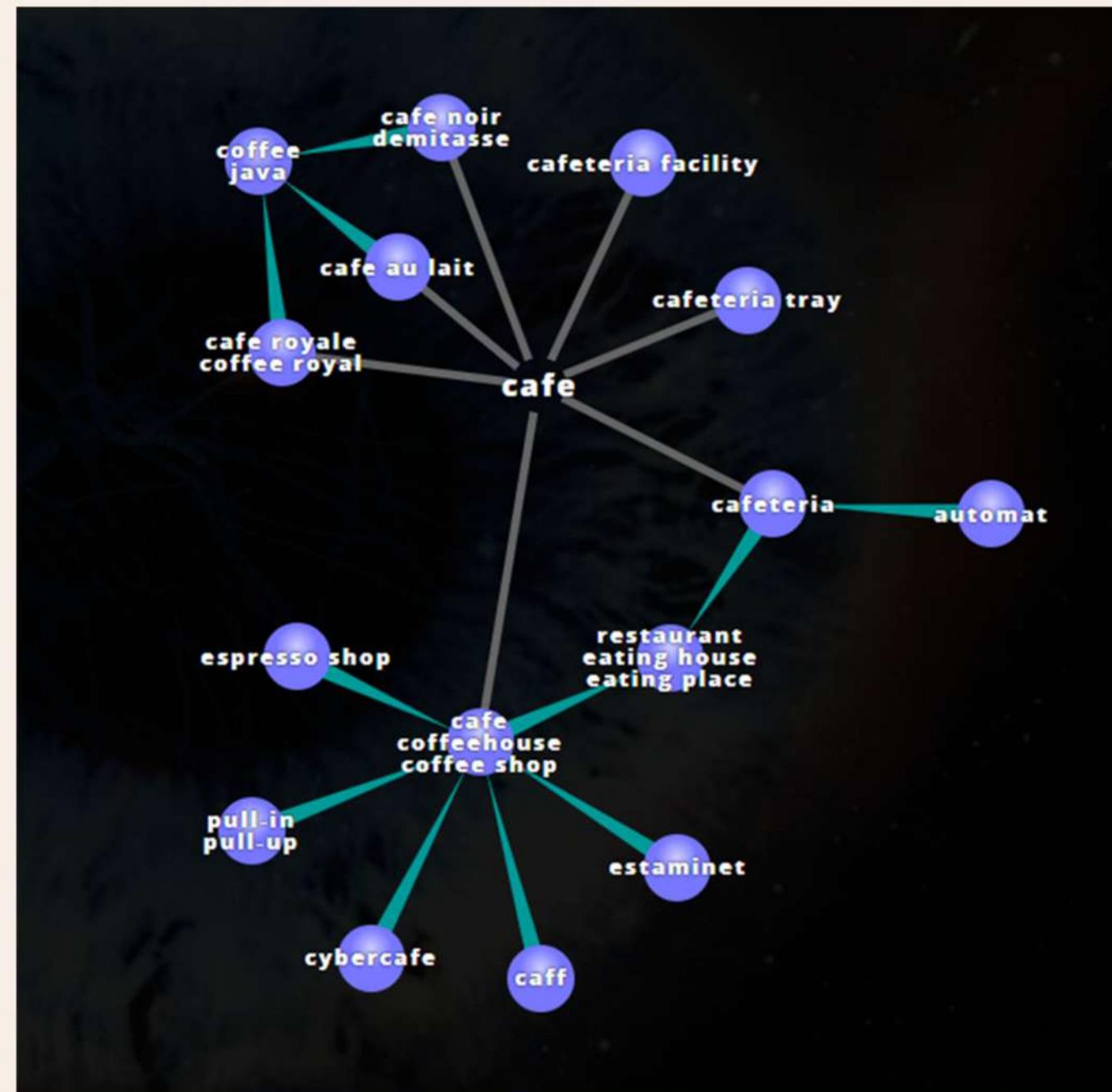


Marzena Halama

WARSTWA WYJŚCIOWA - ANALIZA WYNIKU - PRAWDOPODOBIEŃSTWO KLAS

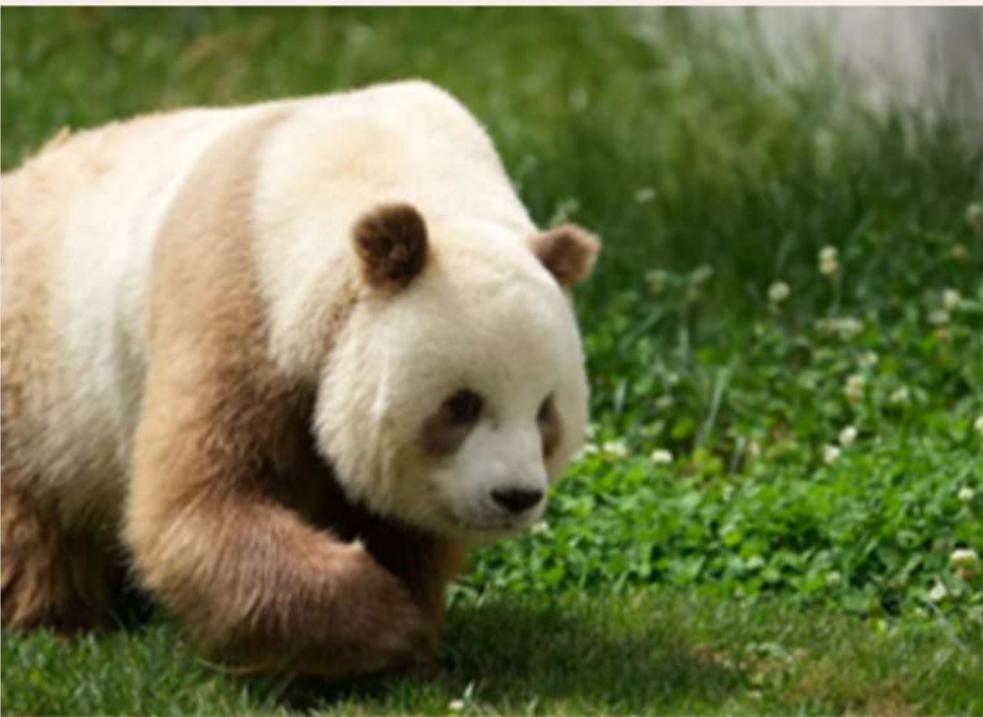


WARSTWA WYJŚCIOWA - ANALIZA WYNIKU - PRAWDOPODOBIEŃSTWO KLAS

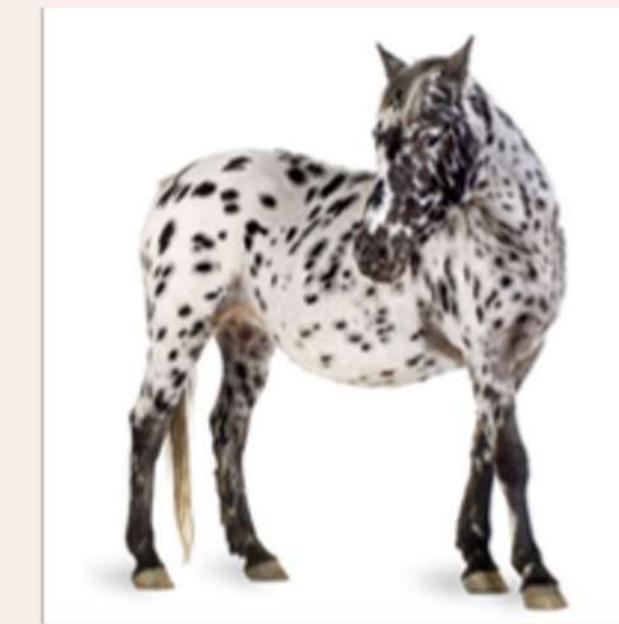


Marzena Halama

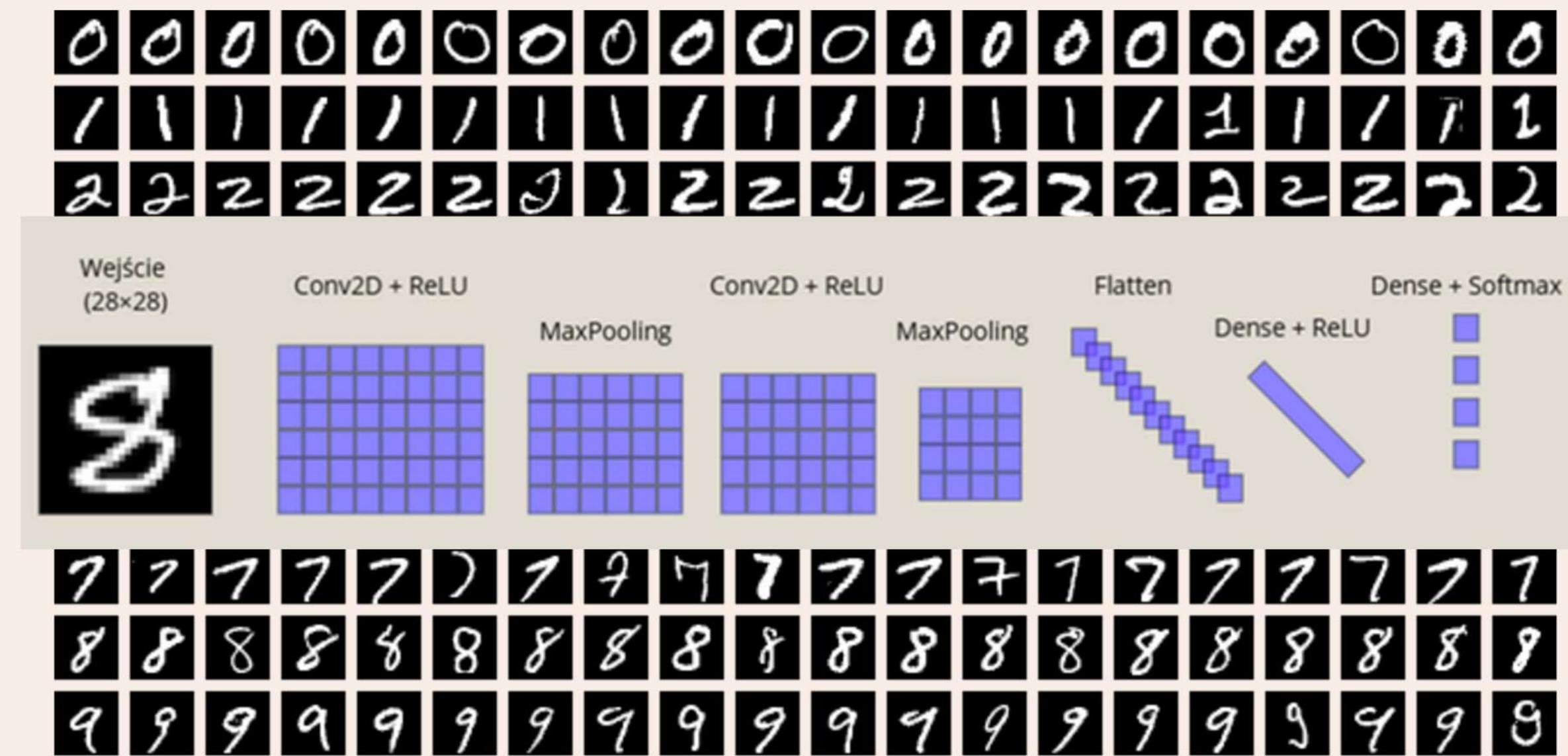
WARSTWA WYJŚCIOWA - ANALIZA WYNIKU - PRAWDOPODOBIEŃSTWO KLAS



Marzena Halama



PIERWSZA WŁASNA SIEĆ NEURONOWA



Marzena Halama

KTÓRE FRAGMENTY OBRAZU WPŁYWAJĄ NA DECYZJĘ SIECI NEURONOWEJ?

01

OBLICZANE SĄ GRADIENTY Z WARSTWY KONWOLUCYJNEJ

02

GRADIENTY TE WSKAZUJĄ, JAK SILNIE KAŻDY NEURON WPŁYWA NA WYNIK
KLASYFIKACJI

03

NA TEJ PODSTAWIE TWORZY SIĘ HEATMAP – POKAZUJE, GDZIE SIEĆ „PATRZYŁA”
PRZY PODEJMOWANIU DECYZJI.

Marzena Halama

GRAD-CAM (GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING)



TOP-5 PREDYKCJE

1. BEAGLE	86.75%
2. WALKER_HOUND	5.32%
3. BLUETICK	3.42%
4. ENGLISH_FOXHOUND	2.67%
5. BASSET	1.52%

Marzena Halama

GRAD-CAM (GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING)

MobileNetV2 - mapa Grad-CAM



ResNet50 - mapa Grad-CAM

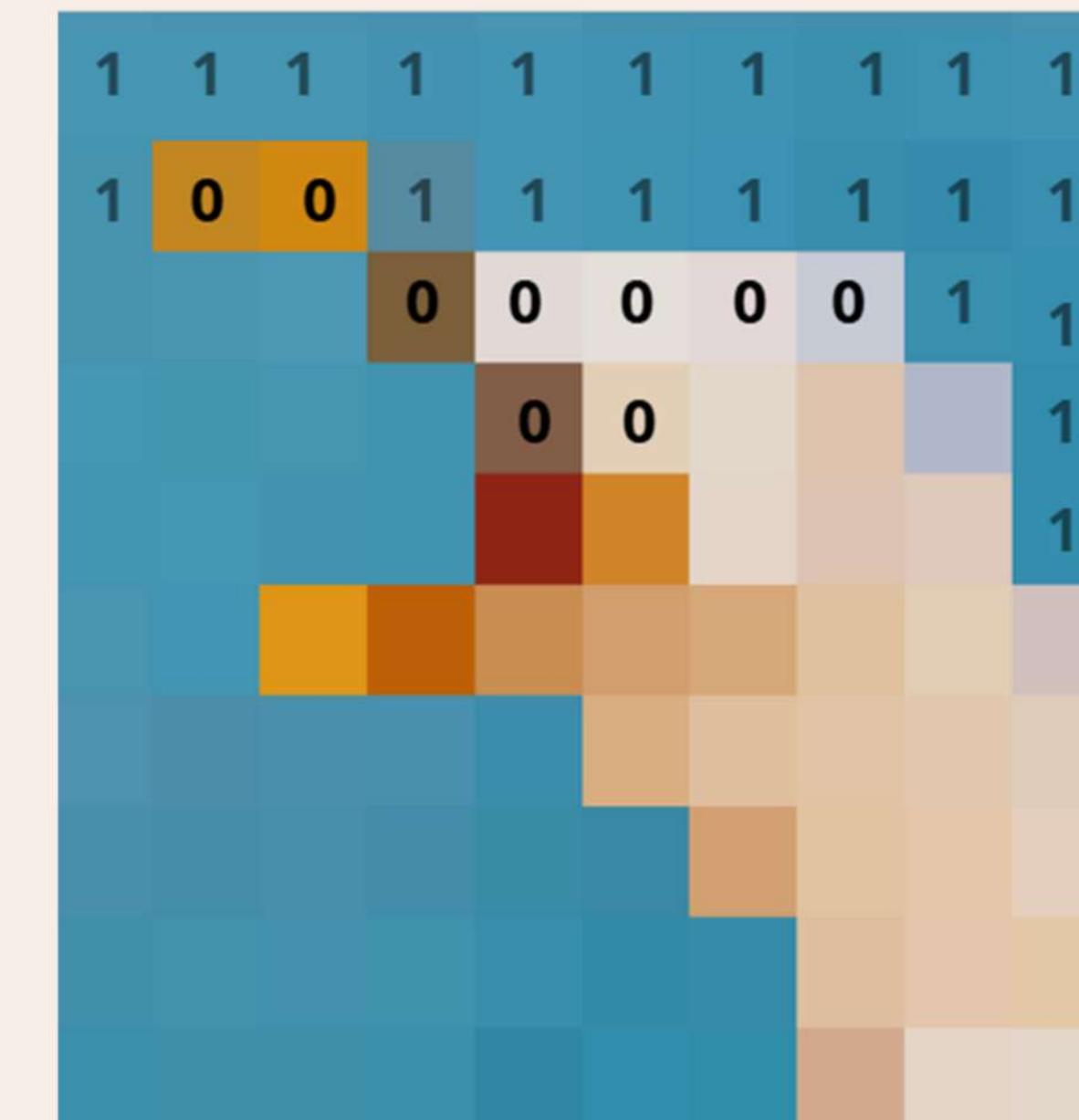
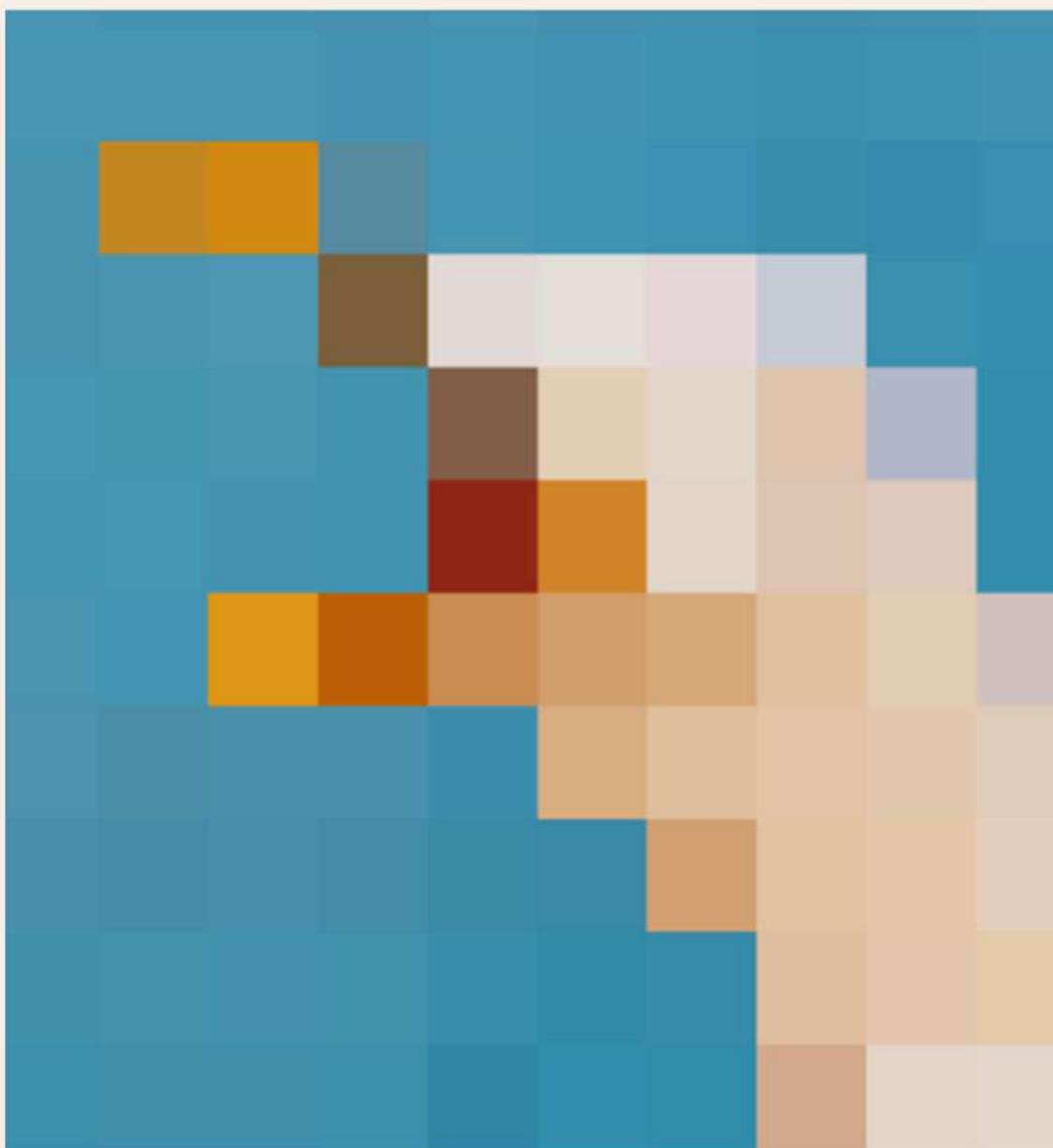


VGG16 - mapa Grad-CAM



Marzena Halama

SEGMENTACJA SEMANTYCZNA SCHEMAT DZIAŁANIA



Marzena Halama

SEGMENTACJA INSTANCJI SCHEMAT DZIAŁANIA



Marzena Halama

WYKORZYSTANIE SEGMENTACJI W PRAKTYCE



Marzena Halama

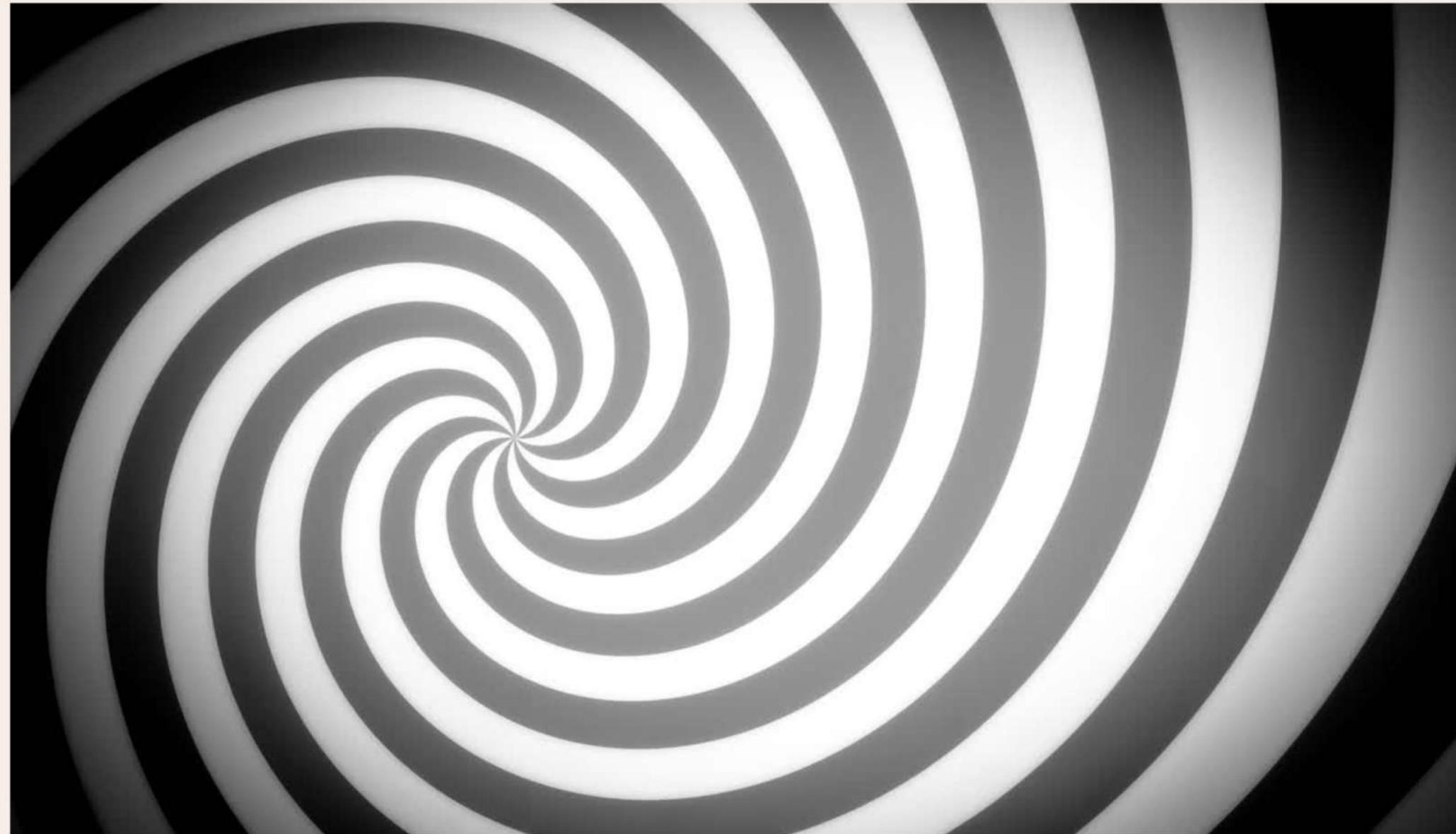
WYZAWANIA ZWIĄZANE Z ANALIZĄ OBRAZÓW

Marzena Halama

Augmentacja danych (ang. data augmentation) to proces sztucznego powiększania zbioru danych treningowych poprzez tworzenie zmodyfikowanych kopii istniejących obrazów.

Celem jest zwiększenie różnorodności danych i poprawa zdolności generalizacji modeli uczących się, np. sieci neuronowych.

WYZAWANIA ZWIĄZANE Z ANALIZĄ OBRAZÓW



MANIPULACJI DZIAŁANIA SIECI
Marzena Halama

CO TO SĄ ATAKI ADWERSARIALNE?

atak adwersarialny to celowe wprowadzenie drobnych, często niemal niewidocznych zmian do danych wejściowych (obrazów, tekstu, sygnału), które powodują, że model ML popełnia błąd (np. błędna klasyfikacja). Atakujący optymalizuje zaburzenie, by maksymalnie „zmylić” model przy minimalnej zmianie wejścia.

Marzena Halama

CO TO SĄ ATAKI ADWERSARIALNE?

atak adwersarialny to celowe wprowadzenie drobnych, często niemal niewidocznych zmian do danych wejściowych (obrazów, tekstu, sygnału), które powodują, że model ML popełnia błąd (np. błędna klasyfikacja). Atakujący optymalizuje zaburzenie, by maksymalnie „zmylić” model przy minimalnej zmianie wejścia.



PANDA 86%

+ 0.007 ×



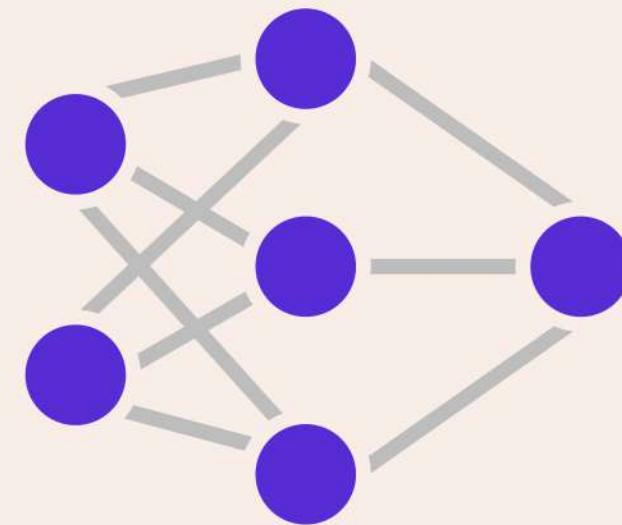
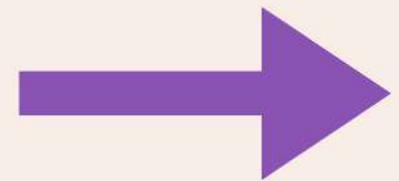
SZUMY

Marzena Halama

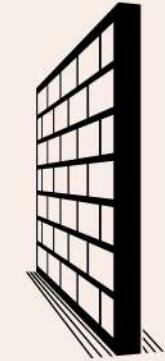


AWOKADO 92%

ORYGINALNY OBRAZ



PORWANA

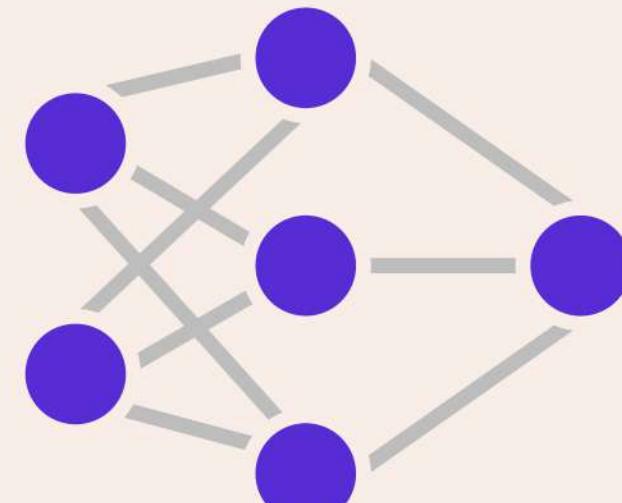
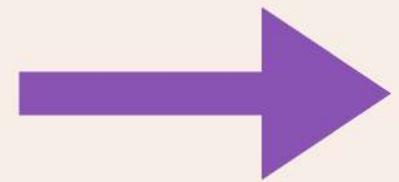


ADWERSARIALNY SZUM



KLASYFIKACJA

ADWERSARIALNY OBRAZ



FAŁSZYWA



Marzena Halama



Repozytorium

SCAN ME



Moja strona

SCAN ME



The graphic features two women against a background of orange autumn leaves and binary code. On the left, a woman in a grey hoodie is laughing. On the right, another woman in a cable-knit sweater with leaf patterns looks towards the camera. A laptop screen in the center displays a grid of small icons. The text "Jesień Pełna Danych" is prominently displayed in large white letters, with "nerds coding gang" written below it in smaller white text.

Marzena Halama



Dziękuje

Marzena Halama

Instytut Informatyki
Teoretycznej i Stosowanej
Polskiej Akademii Nauk

IITiS

PAN
POLSKA AKADEMIA NAUK