Autor: Jakub Tomaszewski
Afiliacja: Politechnika Lodzka
E-mail: 242555@edu.p.lodz.pl

Wyjaśnialność modeli sieci neuronowych z użyciem Captum w porównaniu do danych strukturalnych i obrazowych

Streszczenie

W pracy przedstawiono analizę wyjaśnialności ośmiu modeli sieci neuronowych wytrenowanych na klasycznych zbiorach danych (Iris, Wine, Breast Cancer Wisconsin, MNIST, Imagenette) z wykorzystaniem narzędzi Captum dla ekosystemu PyTorch. Zastosowano metody Integrated Gradients, Saliency, GradientShap oraz Occlusion do generowania lokalnych wyjaśnień decyzji modeli. Lokalnie analizowane przykłady zostały wybrane na podstawie ich odzwierciedlenia wyników globalnych, co pozwoliło uniknąć prezentacji pojedynczych, przypadkowych przypadków, które mogłyby nie reprezentować ogólnej wydajności modelu. Dzięki temu analiza lokalna była spójna z globalnym zachowaniem sieci, co umożliwiło rzetelne identyfikowanie kluczowych cech i formułowanie uogólnionych wniosków dotyczących działania modeli. Wyniki eksperymentów pokazują, że wybrane techniki pozwalają na identyfikację cech wpływających na predykcje, a także wskazują ograniczenia i wyzwania związane z interpretacją modeli głębokich, zwłaszcza w zadaniach przetwarzania obrazów.

Wprowadzenie

Sztuczne sieci neuronowe (ang. Artificial Neural Networks, ANN) stanowią obecnie fundament wielu nowoczesnych rozwiązań uczenia maszynowego, osiągając znakomite wyniki w zadaniach klasyfikacji, regresji oraz rozpoznawania obrazów. Sieć neuronową można formalnie opisać jako funkcję:

$$f_\theta : X \to Y$$

gdzie:

- X — przestrzeń danych wejściowych (np. obrazy, dane tabelaryczne),
- Y — przestrzeń danych wyjściowych (np. etykiety klas),
- θ — parametry modelu, które są optymalizowane podczas procesu uczenia.

Proces uczenia polega na znalezieniu optymalnych parametrów θ na podstawie zbioru treningowego:

$$\{(x_i, y_i)\}_{i=1}^{N}, \quad x_i \in X, \quad y_i \in Y$$

Mimo sukcesów w zakresie dokładności predykcji, złożoność i nieliniowość sieci prowadzi do problemu tzw. „czarnej skrzynki” — modelu, którego wewnętrzne mechanizmy podejmowania decyzji pozostają nieprzejrzyste i trudne do interpretacji. W wielu dziedzinach, w tym w medycynie, finansach czy

systemach rekomendacyjnych, brak transparentności ogranicza zaufanie użytkowników i utrudnia wdrażanie rozwiązań opartych na głębokim uczeniu.

W odpowiedzi na te wyzwania rozwijane są techniki objaśnialnej sztucznej inteligencji (Explainable AI, XAI), które mają na celu dostarczenie interpretowalnych i zrozumiałych wyjaśnień dla predykcji modeli. W szczególności wyróżnia się metody lokalne, które analizują indywidualne przykłady (x_i, y_i) w celu określenia, które cechy wejściowe miały największy wpływ na decyzję modelu $f_\theta(x_i)$. Takie lokalne wyjaśnienia mogą być następnie agregowane i analizowane globalnie, aby sformułować wnioski o ogólnym działaniu modelu.

Celem niniejszej pracy jest kompleksowa analiza i porównanie skuteczności wybranych metod XAI — Integrated Gradients, Saliency, GradientShap oraz Occlusion — w wyjaśnianiu działania różnych architektur sieci neuronowych (w tym MLP i CNN) wytrenowanych na klasycznych zbiorach danych o zróżnicowanej naturze (tabelaryczne, obrazy). W badaniu szczególną uwagę poświęcono doborowi lokalnych przykładów, które są reprezentatywne z perspektywy wyników globalnych modeli, co pozwala na rzetelne i spójne wnioskowanie na temat zachowania sieci.

Powiązane prace

Wyjaśnialność modeli uczenia maszynowego jest kluczowym zagadnieniem, szczególnie w kontekście złożonych modeli, takich jak głębokie sieci neuronowe. W literaturze metody wyjaśniania działania modeli dzieli się na dwie główne klasy: modele samo-interpretowalne (ang. intrinsically interpretable models) oraz metody post-hoc, które stosuje się do analizy modeli o wysokiej złożoności, niedających się łatwo zinterpretować bezpośrednio.

Modele samo-interpretowalne

Modele takie jak drzewa decyzyjne, regresja liniowa czy reguły indukcyjne mają strukturę, która umożliwia bezpośrednie zrozumienie, jakie cechy i w jaki sposób wpływają na predykcję. Pomimo swojej interpretowalności często cechują się ograniczoną elastycznością i mniejszą skutecznością na złożonych danych[1].

Metody post-hoc

W przypadku sieci neuronowych oraz innych modeli o dużej złożoności, stosuje się metody post-hoc, które generują wyjaśnienia po dokonaniu predykcji przez model. Do najczęściej stosowanych podejść należą:

- **Metody atrybucji cech** — przypisują każdej cechce wejściowej wartość wskazującą na jej wpływ na decyzję modelu dla konkretnego przykładu. Należą do nich:
    - **Saliency Maps**[2], które wykorzystują gradienty względem wejścia do wskazania ważnych obszarów.
    - **Integrated Gradients (IG)**[3], które integrują gradienty na ścieżce od wejścia bazowego do obserwowanego przykładu, zapewniając spełnienie pożądanych aksjomatów interpretowalności.
    - **GradientShap**[4], łączące gradienty z metodą Shapley'ów.
    - **Occlusion**[5], polegająca na systematycznym zasłanianiu części wejścia w celu oceny wpływu na wynik predykcji.

    - Metody takie jak **LIME**[4] oraz **SHAP**[6], które są popularne również w analizie danych tabelarycznych i obrazów.

- **Metody kontrfaktyczne** — polegają na znajdowaniu minimalnych zmian wejścia, które zmieniają decyzję modelu, co pozwala zrozumieć granice klasyfikacji[7].
- **Metody globalne** — budują uproszczone, interpretowalne modele zastępcze (np. drzewa decyzyjne, regresję liniową) na podstawie zbioru predykcji modelu złożonego, lub analizują rozkłady atrybucji cech w całym zbiorze danych, aby wyciągnąć ogólne wnioski dotyczące działania modelu[6,8].

Metody

W badaniu przeanalizowano osiem modeli sieci neuronowych różnych architektur i na różnych zbiorach danych, wykorzystując cztery główne metody wyjaśnialności oparte na narzędziu Captum dla ekosystemu PyTorch. Metody te służą do generowania lokalnych wyjaśnień predykcji modeli, które pozwalają na ocenę, które cechy wejściowe (np. piksele obrazu lub cechy tabelaryczne) miały największy wpływ na wynik modelu. Poniżej opisano każdą z zastosowanych technik.

Integrated Gradients (IG)

Integrated Gradients to metoda atrybucji cech, która polega na całkowaniu gradientów modelu względem wejścia wzdłuż ścieżki od tzw. punktu referencyjnego (baseline), będącego wartością reprezentującą "neutralne" wejście, do badanego przykładu. Dzięki temu podejściu możliwe jest uwzględnienie całej trajektorii pomiędzy punktem odniesienia a przykładem, co redukuje szum i daje bardziej stabilne i interpretowalne wyniki. Metoda ta nie wymaga modyfikacji architektury modelu i jest szeroko stosowana zarówno dla danych obrazowych, jak i tabelarycznych.

Saliency Maps (Mapa istotności)

Metoda Saliency opiera się na wyliczeniu gradientu wyjścia modelu względem wejścia. Gradient wskazuje, jak niewielka zmiana wartości wejściowej cechy wpłynie na zmianę wyniku predykcji. Cechy z najwyższymi wartościami gradientów mają największy wpływ na decyzję modelu. W przypadku obrazów gradienty te wizualizowane są jako mapy cieplne, które pokazują najważniejsze piksele. Metoda jest szybka i łatwa do zastosowania, ale może być podatna na szum i dawać mniej stabilne wyjaśnienia niż metody integracyjne.

GradientShap

GradientShap łączy zalety atrybucji gradientowych oraz teorii wartości Shapleya. Bazuje na losowym próbkowaniu punktów referencyjnych wokół baseline i obliczaniu średniej atrybucji gradientowej dla tych punktów. Metoda ta jest bardziej odporna na szum i daje stabilniejsze wyjaśnienia, szczególnie w modelach o wysokiej niestabilności gradientów.

Occlusion (Zasłanianie)

Metoda Occlusion polega na systematycznym "zasłanianiu" fragmentów wejścia i obserwowaniu wpływu tych zmian na wynik modelu. Dla danych obrazowych oznacza to np. zastępowanie fragmentów obrazu (małych okienek pikseli) wartością neutralną i mierzenie różnicy w predykcji. Obszary, których zasłonięcie powoduje największą zmianę, uznawane są za kluczowe dla decyzji modelu. Metoda jest intuicyjna, ale kosztowna obliczeniowo, zwłaszcza dla dużych obrazów. Wizualizacje odbywają się w formie map cieplnych lub wykresów słupkowych.

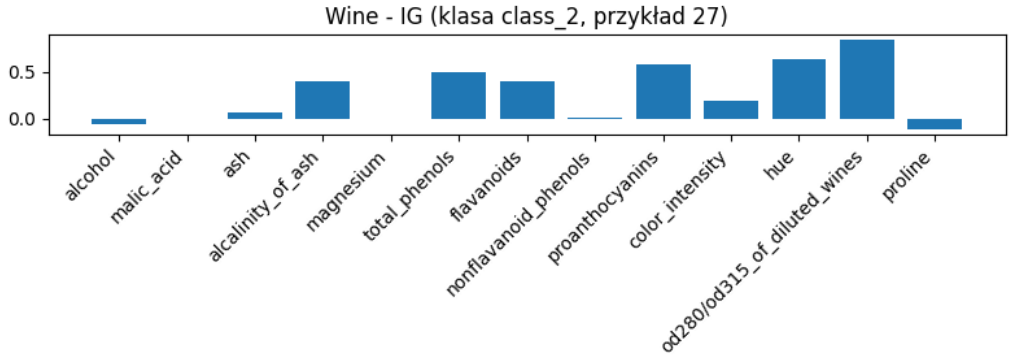
Wyniki

Modele MLP

Wine

Zgodnie z wiedzą dziedzinową, najważniejsze cechy dla klasyfikacji win to 'alcohol', 'flavanoids' oraz 'proline'. Analiza modelu potwierdziła, że dla pierwszych dwóch klas (Barolo i Barbera) rzeczywiście kluczowe były cechy związane z zawartością alkoholu i obecnością flawonoidów. Jednak przy przypisywaniu obiektów do trzeciej klasy (Grignolino) model, oprócz 'flavanoids', wykorzystywał również cechy 'color intensity' oraz 'OD280/OD315 of diluted wines', które okazały się istotne dla rozróżnienia tego szczepu od pozostałych. Przyklad poprawnego przypisania do klasy Gringnolino, nie kierujac sie cechami proline oraz alkohol.

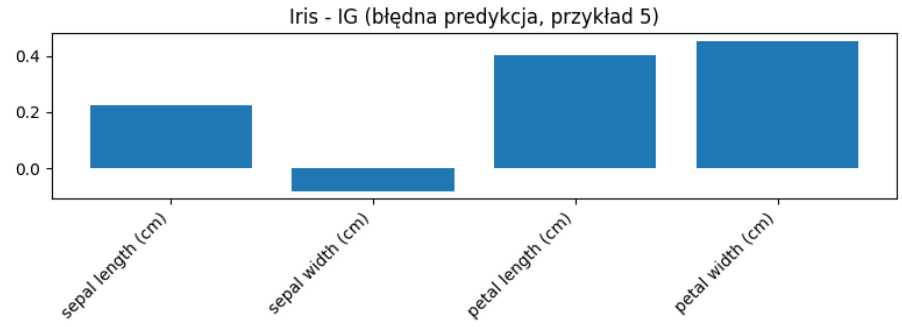
```
Przykład 27: Predykcja: class_2, Pewność: 0.964
Atrybucje IG: [-0.066 -0.009  0.061  0.401 -0.004  0.489  0.4    0.003  0.585  0.186
  0.639  0.846 -0.119]
```



Iris

Wyniki pokazują, że cechy takie jak długość i szerokość płatków (petal length i petal width) w zbiorze Iris mają największy wpływ na decyzję modelu, co jest zgodne z wiedzą ekspercką. Model popełnia jednak niewielką liczbę błędnych predykcji, często myląc klasę versicolor z virginica. W tych przypadkach cecha sepal width zdaje się mieć ograniczone znaczenie dla modelu, co może wskazywać na jej mniejszą rolę w rozróżnianiu tych dwóch klas

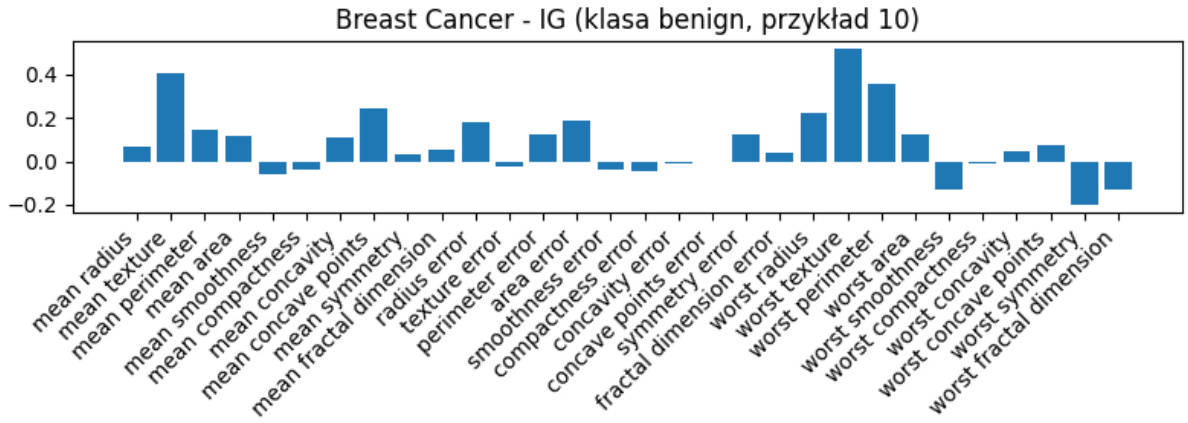
```
Przykłady błędnych predykcji w zbiorze Iris:
Przykład 5: Prawdziwa klasa: versicolor, Predykcja: virginica, Pewność: 0.561
Atrybucje IG: [ 0.225 -0.081  0.403  0.453]
```



Breast Cancer

Model mocniej uwzględnia cechy związane z rozmiarem i kształtem (radius, perimeter, concavity), które faktycznie są kluczowe do rozróżnienia łagodnych i złośliwych guzów.Zauważalne jest także skupienie na

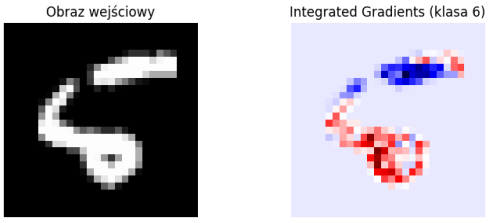
cechach „worst”, które w praktyce dobrze opisują agresywność guza. Cechy typu smoothness, symmetry i fractal dimension mają niższe wartości, co także jest zgodne z tym, że są mniej decydujące.



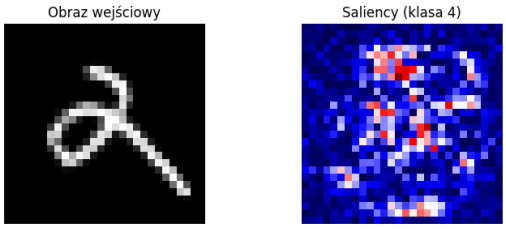
Mnist

MLP 784 sugerowal sie wyraznymi obszarami czesto o centralnej czesci obrazkow, tam gdzie najczesciej biale piksele wystepuja lub nie wystepuja. Kilka przykladow które zostaly blednie sklasyfikowane wynika z niestarannosci w zapisaniu liczby. Te liczby naprawde sa nietypowo zapisane i nie znajac ludzkich schamatow pisania (linia po linii) – latwo zrobic blad:

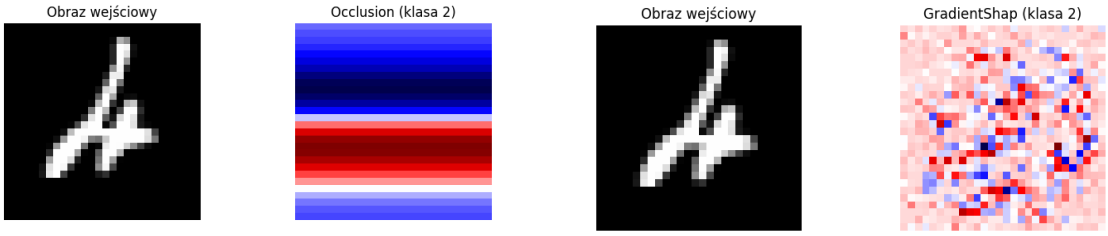
Liczba 5 sklasyfikowana jako 6.



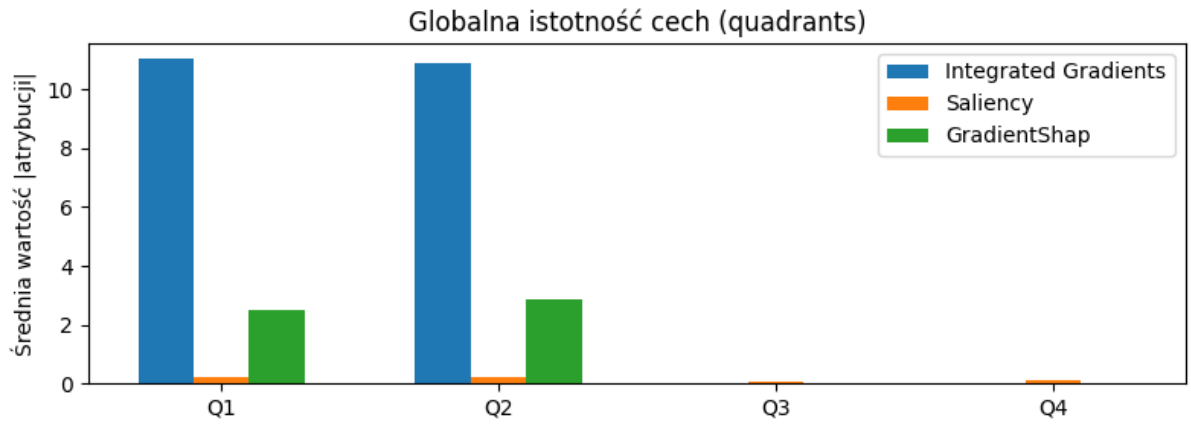
Liczba sklasyfikowana 2 jako 4.



Liczba 4 sklasyfikowana jako 2



Dla modelu Mnist Quadrants ktory osiagnal dosyc niski wynki acuracy=0.39 na zbiorze testowym. Mozna latwo zauwazyc ze model kierowal sie glownie dwoma pierwszymi cwiartkami Q1 i Q2, czyli górna czescia obrazkow.

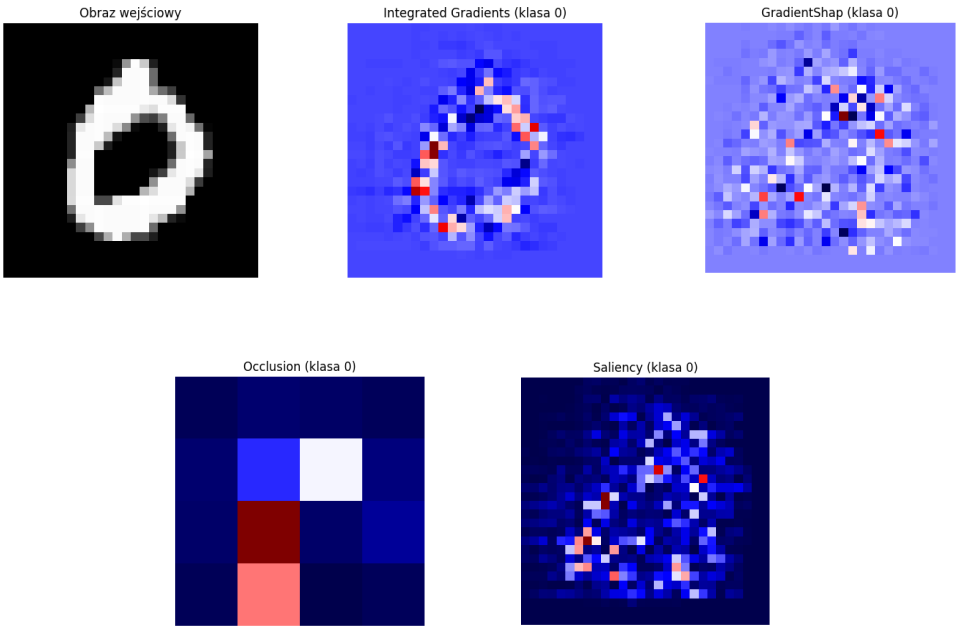


W przypadku PCA które osiagnelo accuracy = 0.482 trudno ocenic istotnosc poszczegolnych cech i znalezc jakies zwiazki ze wzgledny na dwa podstawowe komponenty, ktore globalnie byly wazne w bardzo podobnym stopniu.

Modele CNN

Dla zbioru MNIST

Model CNN mnist zachowal podobna skutecznosc co model Mnist MLP 784. Oto przyklad skuteczniej klasyfikacji:



Model CNN trenowany na zbiorze MNIST zachowuje podobną skuteczność klasyfikacji jak model MLP 784. Na załączonym przykładzie możemy zobaczyć, jak model CNN rozpoznaje cyfrę „0” dzięki różnym metodom interpretacyjnym.

We wszystkich metodach wizualizacji (Integrated Gradients, GradientShap, Occlusion, Saliency) wyraźnie widać, że model skupił się przede wszystkim na kształcie i ciągłości obwodu cyfry „0”. Największe znaczenie mają piksele tworzące kontur pętli oraz kluczowe fragmenty takie jak góra, dół i boki obiektu.

Co ważne, model nie opiera swojej decyzji na pojedynczych pikselach, które mogłyby być przypadkowe czy zaszumione, lecz na całym wzorze i jego strukturze, co świadczy o solidnym i głębokim rozumieniu.

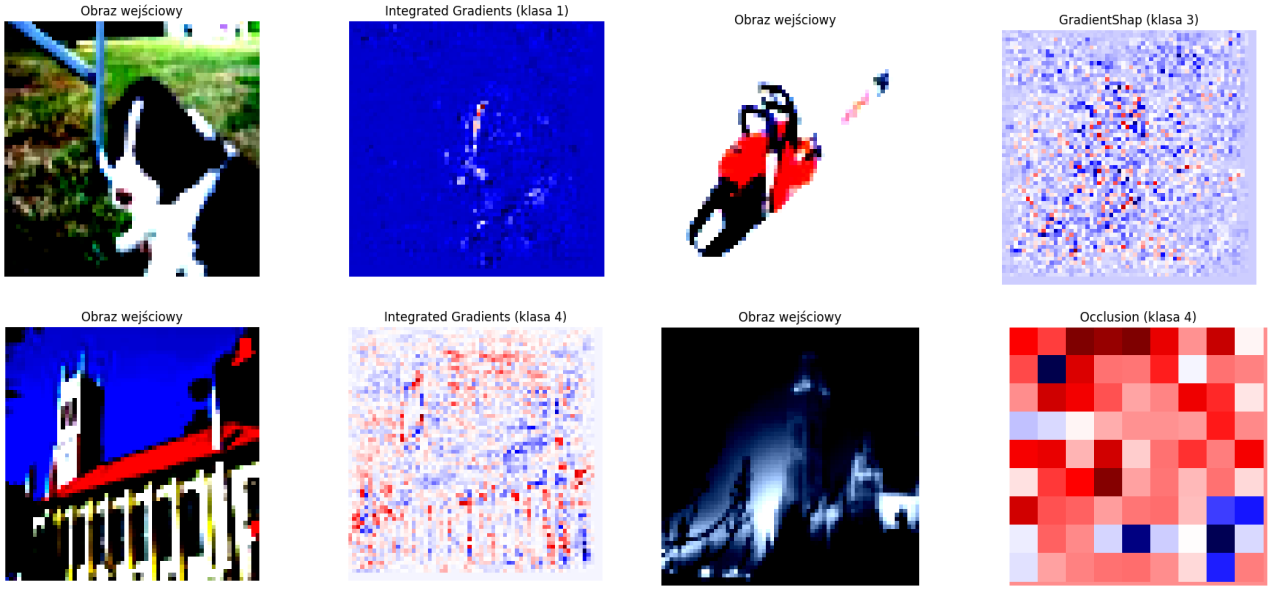
Dla zbioru Imagenette

Model osiagnal skutecznosc accuracy na poziomie okolo 60 %.

Oto tabela przedstawiajaca skutecznosc w zaleznosci od szukanego obiektu:

Klasa	Skuteczność (poprawne/całkowite)
tench	0.734 (284/387)
English springer	0.666 (263/395)
cassette player	0.697 (249/357)
chain saw	0.311 (120/386)
church	0.645 (264/409)
French horn	0.533 (210/394)
garbage truck	0.622 (242/389)
gas pump	0.480 (201/419)
golf ball	0.617 (246/399)
parachute	0.608 (237/390)

Najlepiej klasyfikowane byly ryba trench, english springer i cassete player.



Model CNN rozpoznawał obrazy, skupiając się na charakterystycznych cechach odpowiadających poszczególnym klasom. Metoda **Integrated Gradients** ujawniła, że dla klasy 1 model koncentrował się na precyzyjnych fragmentach obrazu, które silnie wpływały na decyzję, choć sygnał był dość rozproszony i słabo widoczny (np. kontury obiektu). W przypadku klasy 4, Integrated Gradients wykazała bardziej zróżnicowane obszary o różnej sile wpływu, co sugeruje, że model korzysta z wielu cech jednocześnie. Metoda **GradientShap** pokazała bardziej rozproszoną i delikatną strukturę wpływu, co może wskazywać na większą stabilność interpretacji, choć czasem trudną do jednoznacznej interpretacji wizualnej. Z kolei **Occlusion** ujawniła kluczowe fragmenty obrazu o istotnym wpływie na klasyfikację, podkreślając obszary, których zasłonięcie wyraźnie obniżało pewność modelu, co dobrze ilustruje istotność przestrzennych cech w decyzji sieci.

Podsumowanie

W pracy przedstawiono kompleksową analizę wyjaśnialności działania ośmiu modeli sieci neuronowych różnych architektur i na zróżnicowanych zbiorach danych, z wykorzystaniem narzędzi Captum i metod: Integrated Gradients, Saliency, GradientShap oraz Occlusion. Zastosowanie lokalnych wyjaśnień w połączeniu z ich analizą globalną pozwoliło na zidentyfikowanie kluczowych cech wpływających na decyzje modeli oraz na potwierdzenie zgodności wyników z wiedzą ekspercką. Jednocześnie badanie wykazało ograniczenia metod XAI, zwłaszcza przy trudnych zadaniach obrazowych, gdzie interpretacja może być utrudniona przez szum i wysoką złożoność cech.

Dobór reprezentatywnych przykładów lokalnych umożliwił uniknięcie błędów interpretacyjnych wynikających z jednostkowych, niereprezentatywnych przypadków. Praca stanowi wkład w rozwój narzędzi do interpretacji sieci neuronowych i wskazuje kierunki dalszych badań, m.in. w zakresie standaryzacji wyboru przykładów oraz integracji wyjaśnień lokalnych i globalnych.

Spis Literatury

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
[2] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034.
[3] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. Proceedings of ICML 2017.
[4] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NIPS).
[5] Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. European Conference on Computer Vision (ECCV).
[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of KDD 2016.
[7] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology.
[8] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. Proceedings of KDD 2015.