Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Informatyka techniczna**

PRACA DYPLOMOWA
MAGISTERSKA

Porównanie metod XAI w klasyfikacji obrazów

inż. Mateusz Moroszczuk

Opiekun pracy
dr hab. inż. Henryk Maciejewski

Słowa kluczowe: XAI, klasyfikacja obrazów

WROCŁAW 2024

STRESZCZENIE

ABSTRACT

SPIS TREŚCI

Wprowadznie . . . 3
Wprowadzenie do problemu . . . 3
Cel pracy . . . 3
Zakres pracy . . . 3
Wybrane Metody XAI . . . 3
Przegląd literatury . . . 5
Literatura dotycząca metod XAI . . . 5
Literatura dotycząca metod klasyfikacji obrazów . . . 5
Metodologia . . . 6
Przygotowanie środowiska . . . 6
Charakterystyka wybranych metod XAI . . . 6
LIME . . . 6
SHAP . . . 7
Grad-CAM . . . 7
Wybór hiperparametrów . . . 7
Algorytmy i miary jakości . . . 7
Eksperymenty i wyniki . . . 9
Analiza porównawcza spójności wyjaśnień . . . 9
Analiza porównacza wyjaśnień . . . 9
Łączenie wyjaśnień różnych metod . . . 9
Porównanie z innymi modelami klasyfikacji obrazów . . . 9
Ocena poszczególnych algorytmów . . . 9
Dyskusja . . . 10
Omówienie wyników . . . 10
Wnioski . . . 10
Wnioski . . . 11
Podsumowanie . . . 12
Spis rysunków . . . 13
Spis listingów . . . 14
Spis tabel . . . 15
Dodatki . . . 16

A. Dodatek 1 17

WPROWADZNIE

WPROWADZENIE DO PROBLEMU

Przedstawienie problemu klasyfikacji obrazów przy użyciu sieci głębokich.
Potrzeba XAI do zrozumienia podejmowanych decyzji.

CEL PRACY

Przedstawienie celu: zbadanie i porównanie metod XAI. O zbiorze ImageNet, skupić się na klasach obrazów charakteryzujących się niską 'robustness'

ZAKRES PRACY

Przygotowanie środowiska programistycznego i zbioru danych. Analiza lieratury Badania różnych metod XAI

WYBRANE METODY XAI

W dziedzinie wyjaśnialnej sztucznej inteligencji (XAI) istnieje wiele metod służących do wyjaśnienia decyzji podejmowanych przez modele uczenia maszynowego, zwłaszcza sieci neuronowe. W naszej pracy skupiamy się na trzech głównych metodach XAI, które są szeroko stosowane i dobrze zbadane:

1. **LIME (Local Interpretable Model-agnostic Explanations)**: LIME jest techniką stosowaną do generowania lokalnych interpretacji modeli uczenia maszynowego, które są agnostyczne względem modelu. Metoda ta polega na generowaniu lokalnych wyjaśnień dla indywidualnych predykcji modelu, co pozwala na zrozumienie, dlaczego model dokonał konkretnej klasyfikacji dla danego przypadku testowego.
2. **SHAP (SHapley Additive exPlanations)**: SHAP to metoda oparta na teorii gier, która dostarcza globalnych interpretacji modeli uczenia maszynowego. Wykorzystuje ona wartości Shapleya, aby obliczyć wpływ każdej cechy na predykcje modelu. SHAP umożliwia zrozumienie, jak poszczególne cechy przyczyniają się do wyników modelu na poziomie globalnym.
3. **GradCAM (Gradient-weighted Class Activation Mapping)**: GradCAM to technika wizualizacji, która pozwala na lokalizowanie istotnych obszarów na obrazie, które

przyczyniają się do konkretnej predykcji modelu. Wykorzystuje ona gradienty ostatniej warstwy sieci neuronowej w celu generowania map aktywacji klas, co umożliwia zrozumienie, które obszary obrazu były najistotniejsze dla decyzji modelu.

PRZEGLĄD LITERATURY

LITERATURA DOTYCZĄCA METOD XAI

Przegląd literatury dotyczącej XAI. Podsumowanie aktualnych badań na temat tych metod, ich zastosowania, a także porównania wnioski wyciągnięte przez badaczy. Aktualne wyzwania związane z wyjaśnianiem wyników modeli głębokiego uczenia

LITERATURA DOTYCZĄCA METOD KLASYFIKACJI OBRAZÓW

Przegląd literatury dotyczącej klasyfikacji obrazów. Jak zostały zbudowane. i jakie metody użyte do szkolenia i oceny. Wyzwania i zagrożenia np. 'robustness'

METODOLOGIA

PRZYGOTOWANIE ŚRODOWISKA

Do przeprowadzenia eksperymentów z wybranymi metodami XAI wymagane jest odpowiednie przygotowanie środowiska programistycznego. W naszej pracy wykorzystujemy język Python oraz kilka popularnych bibliotek do uczenia maszynowego i przetwarzania obrazów. Poniżej przedstawiamy krótki opis wykorzystywanych funkcji i bibliotek:

— **Python** - język programowania, który jest szeroko stosowany w dziedzinie uczenia maszynowego.
— **TensorFlow** - biblioteka do uczenia maszynowego, która jest szeroko stosowana w dziedzinie uczenia maszynowego.
— **scikit-image** - biblioteka do przetwarzania obrazów, która zawiera wiele funkcji do przetwarzania obrazów.
— **LIME** - biblioteka do wyjaśniania modeli uczenia maszynowego, która jest szeroko stosowana w dziedzinie uczenia maszynowego.
— **SHAP** - biblioteka do wyjaśniania modeli uczenia maszynowego, która jest szeroko stosowana w dziedzinie uczenia maszynowego.
— **numpy** - biblioteka do obliczeń numerycznych, która jest szeroko stosowana w dziedzinie uczenia maszynowego.
— **matplotlib** - biblioteka do tworzenia wykresów, która jest szeroko stosowana w dziedzinie uczenia maszynowego.

CHARAKTERYSTYKA WYBRANYCH METOD XAI

W tej sekcji dokładniej omówimy trzy wybrane metody wyjaśnialnej sztucznej inteligencji (XAI), które zostaly wcześniej wymienione: LIME, SHAP oraz Grad-CAM.

LIME

LIME jest techniką wyjaśnialnej sztucznej inteligencji, która generuje lokalne interpretacje modeli uczenia maszynowego. Metoda ta jest agnostyczna względem modelu, co oznacza, że może być stosowana do różnych rodzajów modeli, niezależnie od ich architektury. LIME działa poprzez tworzenie lokalnych modeli, które starają się naśladować oryginalny model dla konkretnego przypadku. Jest to przydatne narzędzie do zrozumienia, dlaczego model dokonał konkretnej klasyfikacji dla danego przypadku testowego.

SHAP

SHAP jest metodą opartą na teorii gier, która dostarcza globalnych interpretacji modeli uczenia maszynowego. Wykorzystuje wartości Shapleya, aby obliczyć wpływ każdej cechy na predykcje modelu. SHAP umożliwia zrozumienie, jak poszczególne cechy przyczyniają się do wyników modelu na poziomie globalnym. Jest to przydatne narzędzie do identyfikowania najważniejszych cech wpływających na predykcje modelu.

Grad-CAM

GradCAM to technika wizualizacji, która pozwala na lokalizowanie istotnych obszarów na obrazie, które przyczyniają się do konkretnej predykcji modelu. Jest to przydatne narzędzie do zrozumienia, które obszary obrazu były decydujące dla decyzji modelu. Grad-CAM wykorzystuje gradienty ostatniej warstwy sieci neuronowej w celu generowania map aktywacji klas, co umożliwia lokalizowanie obszarów najbardziej istotnych dla klasyfikacji.

WYBÓR HIPERPARAMETRÓW

ALGORYTMY I MIARY JAKOŚCI

W tej sekcji opisujemy algorytmy oraz miary jakości użyte do oceny skuteczności metod wyjaśnialnej sztucznej inteligencji (XAI) w zadaniu klasyfikacji obrazów. Możemy je podzielić na dwie kategorie, oparte na prawdzie oraz oparte na modelu.

Miary oparte na prawdzie są to miary, które oceniają skuteczność metod XAI na podstawie faktycznych wartości docelowych.

Natomiast **miary oparte na modelu** oceniają skuteczność metod XAI do wytłumaczania modelu.

1. **IOU** (Intersection over Union): Algorytm IOU jest powszechnie stosowany w zadaniach segmentacji obrazów do oceny jakości detekcji obrazów. Oblicza on stosunek powierzchni przecięcia dwóch obszarów do ich sumy. W naszym kontekście, IOU może być używany do oceny zgodności obszarów wyznaczonych przez metody XAI z rzeczywistymi obiektami na obrazie.
2. **Confidence change**: Miara ta ocenia zmianę pewności klasyfikacji na obrazach po zastosowaniu metod XAI. Jest to różnica między pewnościami klasyfikacji na oryginalnych obrazach a pewnościami klasyfikacji na obrazach zmodyfikowanych z użyciem XAI. Dzięki tej mierze możemy ocenić, czy metody XAI wpływają na zmianę pewności klasyfikacji. Obszar wyznaczony przez metody XAI powinien minimalnie zmniejszyć pewność klasyfikacji lub ją zwiększyć. Natomiast obszar poza wyznaczonym obszarem powinien drastycznie zmniejszyć pewność klasyfikacji.

Poprzez zastosowanie tych algorytmów i miar jakości, będziemy w stanie ocenić skuteczność metod XAI w wyjaśnianiu klasyfikacji obrazów oraz ich wpływ na jakość klasyfikacji.

EKSPERYMENTY I WYNIKI

ANALIZA PORÓWNAWCZA SPÓJNOŚCI WYJAŚNIEŃ

Porównanie spójności wyjaśnień między różnymi metodamii. Porównanie spoójności w zależności od klasy obrazów

ANALIZA PORÓWNACZA WYJAŚNIEŃ

Ocena wyjaśnień IOU, confidence score. Ocena wyjaśnień w zależności od klasy.

ŁĄCZENIE WYJAŚNIEŃ RÓŻNYCH METOD

Porównanie spójności łączonych wyjaśnień (suma, średnia oraz część wspólna) Ocena wyjaśnień dla połączonych wyjaśnień (suma, średnia oraz część wspólna).

PORÓWNANIE Z INNYM ZBIOREM DANYCH

PORÓWNANIE Z INNYMI MODELAMI KLASYFIKACJI OBRAZÓW

OCENA POSZCZEGÓLNYCH ALGORYTMÓW

Ocena poszczególnych algorytmów. Porównanie wyników metod pod kątem ich zdolności do dostarczania zrozumiałych i precyzyjnych wyjaśnień. Słabe i mocne strony każdej metody. Czas wykonania eksperymentów oraz złożoność obliczeniowa

DYSKUSJA

OMÓWIENIE WYNIKÓW

Podsumowanie kluczowych informacji z sekcji "Eksperymenty i wyniki", w kontekście celów pracy. Które metody najskuteczniejsze. Potencjalne przyczyny różnic między metodami i ich wpływ na jakość wyjaśnień

WNIOSKI

WNIOSKI

PODSUMOWANIE

Główne osiągnięcia. Znaczenie pracy. Prespektywa na przyszłość.

SPIS RYSUNKÓW

SPIS LISTINGÓW

SPIS TABEL

Dodatki

A. DODATEK 1

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.