Dr hab. inż. Marek Kowal, prof. UZ
Instytut Sterowania i Systemów Informatycznych
Wydział Informatyki, Elektrotechniki i Automatyki
Uniwersytet Zielonogórski
email: M.Kowal@issi.uz.zgora.pl

Zielona Góra, 27.08.2022r.

Recenzja rozprawy doktorskiej

Tytuł rozprawy: **Methods for Interpretation and Validation of Representations Generated by Deep Convolutional Neural Networks**

Autor rozprawy: **mgr inż. Tomasz Szandała**

Promotor rozprawy: **dr hab. inż. Henryk Maciejewski**

Dziedzina: **nauki techniczne**

Dyscyplina: **informatyka techniczna i telekomunikacja**

1. Cel, zakres i charakter rozprawy

Głównym celem recenzowanej pracy doktorskiej było opracowanie i implementacja nowych bardziej skutecznych metod do interpretacji oraz walidacji wyników uzyskiwanych przez głębokie sieci neuronowe. Metody te ukierunkowane są na wyjaśnienie w przejrzysty i zrozumiały dla człowieka sposób decyzji podejmowanych przez złożone modele neuronowe. Zagadnienia z tego zakresu należą do obszaru badawczego określanego jako wytłumaczalna (interpretowalna) sztuczna inteligencja (ang. Explainable Artificial Intelligence, XAI). Aktualnie jest to bardzo aktywny obszar badawczy. Wynika to z faktu, że głębokie sieci neuronowe są z dużymi sukcesami wykorzystywane między innym w analizie i rozpoznawaniu obrazów oraz w przetwarzaniu języka naturalnego. Stosowane obecnie w praktyce modele neuronowe mają nawet setki warstw neuronów i setki milionów parametrów co oznacza, że musimy je traktować jako tzw. czarne skrzynki. W efekcie walidacja jakości takich modeli często sprowadza się jedynie do wyznaczenia pewnych metryk dla danych testowych. Jednak w przypadku zastosowań w obszarze np. diagnostyki medycznej czy pojazdów autonomicznych jest duża potrzeba poznania i weryfikacji poprawności mechanizmów stojących za decyzjami podejmowanymi przez głębokie sieci neuronowe. Pozwala to zwiększyć bezpieczeństwo rozwiązań opartych o metody uczenia głębokiego. Wspomniane metody wyjaśniające zachowanie głębokich sieci neuronowych są również niezbędne w procesie projektowania i implementowania modeli głębokich sieci neuronowych aby poznawać przyczyny błędnych decyzji modelu i wprowadzać ewentualne korekty. Bez tych metod zdani jesteśmy na technikę prób i błędów, której efektywność jest zwykle niewielka ze względu na ogromną liczbę parametrów i hiper-parametrów w sieciach neuronowych.

Praca swoim zakresem obejmuje przedstawienie obecnego stanu wiedzy w obszarze głębokich sieci neuronowych ze szczególnym uwzględnieniem splotowych sieci neuronowych. Zaprezentowanie stanu wiedzy w obszarze metod wykorzystywanych do interpretacji decyzji głębokich sieci splotowych poprzez analizę atencji sieci neuronowej w wybranych obszarach cech obrazu wejściowego. Rozszerzenie obecnego stanu wiedzy

przez udoskonalenie metod do analizy i interpretacji atencji splotowych sieci neuronowych. Opracowanie miary ilościowej do oceny skupienia sieci neuronowej na pożądanych obiektach do zautomatyzowanej walidacji wielkoskalowych wyników generowanych przez splotowe sieci neuronowe. Wypracowanie kompleksowego mechanizmu oceny skuteczności algorytmów do analizy i interpretacji atencji sieci neuronowych. Opracowanie nowego podejścia do analizy wyników splotowych sieci neuronowych przez wizualizację istotnych cech na obrazach wejściowych i możliwość porównywania aktywności poszczególnych cech między różnymi obrazami wejściowymi. Praca obejmuje również swoim zakresem kompleksowe badania weryfikujące skuteczność opracowanych algorytmów. W ramach tej części przeprowadzono szereg eksperymentów badawczych, dla których zaprezentowano wyniki, przeprowadzono dyskusję i przedstawiono wnioski.
Rozprawa ma charakter doświadczalno-teoretyczny jednak ze znaczenie większym naciskiem na wątek doświadczalny. Taki charakter pracy wynika ze specyfiki poruszanych problemów, które wymagały przeprowadzenia kompleksowych badań empirycznych w celu potwierdzenia ich poprawności i skuteczności. Przeprowadzone rozważania teoretyczne są jedynie elementem pomocniczym pozwalającym zrozumieć ideę zaproponowanych w pracy metod ale nie służą do udowodnienia ich poprawności. Rozprawa ujawnia również charakter aplikacyjny ponieważ opracowane algorytmy zostały zaimplementowane oraz publicznie udostępnione w repozytorium Github.
Uwzględniwszy powyższe fakty uważam, że cel rozprawy Pana magistra inżyniera Tomasza Szandały jest sformułowany właściwie ponieważ dotyczy aktualnego problemu naukowego, którego rozwiązanie ma istotne znaczenie dla rozwoju informatyki technicznej. Zakres pracy uznaję za poprawny, właściwy podjętemu problemowi naukowemu.

2. Zawartość rozprawy

Rozprawę napisano w języku angielskim. Praca składa się ze streszczenia w języku polskim oraz w języku angielskim, 5 rozdziałów oraz spisu literatury. Praca ma 91 stron przy czym część pracy, w której zaprezentowano główne osiągnięcia obejmuje 81 stron (3-83s.).

Rozdział 1 jest wprowadzeniem do tematyki rozprawy. Zaprezentowano w nim przewodnik po rozdziałach rozprawy, motywację pracy, w której zawarto skrótowy opis rozwoju głębokich sieci splotowych oraz przedstawiono zagadnienie interpretowalności głębokich sieci neuronowych. W dalszych podrozdziałach opisano w skrócie wkład jaki wnosi niniejsza praca w rozwój dyscypliny naukowej oraz wskazano repozytorium Github, w którym udostępniono publicznie kody algorytmów, które opracowano w ramach rozprawy.

Rozdział 2 poświęcono na zaprezentowanie aktualnego stanu wiedzy w zakresie splotowych sieci neuronowych oraz metod wizualizacji atencji splotowych sieci neuronowych. W pierwszej części zaprezentowano i opisano między innymi schemat architektury typowej splotowej sieci neuronowej, historię rozwoju splotowych sieci neuronowych oraz typowe udoskonalenia splotowych sieci neuronowych wypracowane na przestrzeni ich rozwoju. Ta część rozdziału składa się tylko z 6 stron i ma charakter poglądowo-przeglądowy. Doktorant nie przedstawia podstaw teoretycznych i szczegółów z zakresu funkcjonowania splotowych sieci neuronowych, prawdopodobnie zakładając, że jest to wiedza już powszechna dla specjalistów z tego zakresu i w związku z tym przedstawia dla poszczególnych zagadnień jedynie odpowiednie referencje literaturowe, które prezentują poruszane temat dogłębniej. W drugiej części rozdziału zaprezentowano

przegląd klasycznych metod wykorzystywanych do wizualizacji i interpretacji atencji splotowych sieci neuronowych. W sumie na 16 stronach zaprezentowano 12 różnych metod. Przygotowane opisy są pobieżne, prezentują tylko ogólne koncepcje prezentowanych metod. Uważam, że praca zyskałaby gdyby chociaż wybrane algorytmy, które posłużyły w dalszej części jaka bazowe do rozwoju autorskich rozwiązań zaprezentowano obszerniej z większymi szczegółami.

Rozdział 3 przedstawia metody wypracowane przez Doktoranta. Składa się z 4 podrozdziałów. W pierwszym zaprezentowano metodę o nazwie *Attention Focus Evaluation*. Metoda ta bazuje na znanym algorytmie GradCAM. Modyfikacja polega na wprowadzeniu automatycznego mechanizmu detekcji obiektów na obrazach za pomocą platformy Detectron2, które następnie służą jako odniesienie do oceny poprawności map atencji wyznaczanych przez GradCAM. Wypracowane rozwiązanie pozwala na ilościową ocenę jakości wyników generowanych przez GradCAM dzięki czemu metoda może być stosowana do zautomatyzowanej analizy wyników w sytuacjach gdy mamy do czynienia z dużą ilością klas i danych. W drugim podpunkcie przedstawiono koncepcję zaproponowanej metody wizualizacji atencji splotowych sieci neuronowych o nazwie *Gradual Extrapolation*. Metoda ta jest odpowiedzią na problem skalowania map atencji do rozmiaru obrazów wejściowych. W klasycznych podejściach jak np. dla algorytmu GradCAM przeskalowanie odbywa się za pomocą zwykłej interpolacji obrazu. W efekcie uzyskane wyniki charakteryzują się brakiem precyzji tzn. wyznaczone mapy wskazują obiekty, na które zwraca uwagę sieć neuronowa ale nie są w stanie sprecyzować, które cechy tych obiektów są istotne. Koncepcję zaproponowanej metody zaprezentowano za pomocą przykładu na rysunku 23. Opracowana metoda może być stosowna z wieloma różnymi metodami wyznaczającymi mapy atencji dla warstw pośrednich. W podrozdziale trzecim zaproponowano metodę o nazwie *Latent Features Detection*. W praktyce jest to złożenie dwóch metod zaprezentowanych w poprzednich podrozdziałach. Podobnie jak w przypadku metody *Attention Focus Evaluation* możliwe jest wyznaczenie ilościowego wskaźnika określającego jakość mapy atencji. Przy czym tym razem dzięki zastosowaniu algorytmu *Gradual Extrapolation* metoda skupia się na cechach klasyfikowanych obiektów a nie na całych obiektach. W podrozdziale 4 zaprezentowano kolejną autorską metodę o nazwie *Principal Image Sections Mapping (PRISM)*. Służy ona do oznaczenia kolorem na obrazie wejściowym cech obrazu, które aktywują w górnych warstwach splotowych (zwykle ostatnich) podobne wysokopoziomowe wzorce. Opracowana metoda należy do klasy podejść tzw. explanation by example czyli wyjaśniania za pomocą przykładów ponieważ pozwala porównywać ze sobą obrazy tych samych lub różnych klas pod kątem aktywacji cech wysokopoziomowych. Sercem metody jest analiza składowych głównych (PCA), która służy do redukcji liczby cech opisujących mapy aktywacji w warstwie splotowej. Uzyskane kolorowe mapy określające poziomy aktywacji cech wysokopoziomowych są następnie propagowane wstecznie aż do uzyskania rozmiaru obrazu wejściowego z wykorzystaniem opracowanego wcześniej algorytmu *Gradual Extrapolation*. Na opisanie wszystkich metod przeznaczono tylko 8 stron. Wprawdzie udało się przekazać ogólne koncepcje poszczególnych metod jednak zabrakło szczegółów, rozważań teoretycznych oraz przykładów, i schematów wyjaśniających działanie tych algorytmów. Cześć brakujących informacji pojawiła się dopiero w rozdziale 4, który poświęcono na opis badań eksperymentalnych. Uważam jednak, że praca zyskałaby gdyby w rozdziale 3 zaprezentowano obszerniej zaproponowane metody.

Rozdział 4 prezentuje wyniki przeprowadzanych badań eksperymentalnych. Rozdział podzielono na dwa główne podrozdziały. W pierwszym zaprezentowano wyniki osiągnięte dla algorytmu *Attention Focus Evaluation* i *Latent Features Detection*. Do przeprowadzenia

eksperymetów dla metody *Attention Focus Evaluation* wykorzystano sieć neuronową ResNet50 oraz VGG a dla algorytmu *Latent Features Detection* tylko sieć neuronową o architekturze VGG16. W obu przypadkach eksperymenty obliczeniowe miały podobny przebieg. Po pierwsze przeprowadzono badanie z wykorzystaniem zbioru ImageNet ze zredukowaną do 9 liczbą klas a następnie powtórzono badanie dla zbioru ImageNet ze wszystkimi klasami. Dla tych eksperymentów zaprezentowano wyniki w formie tabel, w których zamieszczono wskaźniki jakości dla uzyskanych map atencji oraz dokładność klasyfikacji. Ponadto dla każdego scenariusza przeprowadzono szczegółowe studium wybranego przypadku (obrazu) aby zweryfikować skuteczność metod. Zaprezentowano również ograniczenia metod, które zilustrowano przykładami z uzyskanych wyników. Aby potwierdzić skuteczność opracowanych metod w zadaniu wyznaczania atencji sieci neuronowych wykorzystano wiedzę uzyskaną za pomocą opracowanych metod do przeprowadzenia tzw. ataku adwersaryjnego (ang. adversarial attack). Skuteczność przeprowadzonych ataków potwierdziła, że opracowane metody wyznaczają poprawne mapy atencji, wskazujące cechy obiektów, na których sieci neuronowe rzeczywiście skupiają swoją uwagę.

W drugim podrozdziale porównano za pomocą opracowanej techniki *FIAt* metody wizualizacji atencji wykorzystujące algorytm *Gradual Extrapolation* z odpowiednikami, które nie wykorzystują tej metody. Do przeprowadzenia testów porównawczych wykorzystano algorytm GradCAM oraz Excitation Backpropagation. Wyniki porównania za pomocą wskaźnika *FIAt* zaprezentowano w formie tabeli dla klasycznych algorytmów GradCAM i Excitation Backpropagation oraz wersji tych algorytmów wykorzystujących skalowanie wyników za pomocą *Gradual Extrapolation*. Eksperymenty powtórzono dla 5 różnych architektur splotowych sieci neuronowych. Dla wybranych przypadków zaprezentowano szczegółowe wyniki w postaci uzyskanych map atencji. W drugiej części tego podrozdziału skupiono się na przeprowadzeniu badań skuteczności metody *Principal Image Sections Mapping*. Przeprowadzone badania miały charakter studium przypadku. Na przykładzie zweryfikowano, że metoda jest przydatna w identyfikacji precyzyjnych cech dyskryminujących poszczególne klasy obrazów. Porównano wyniki uzyskane za pomocą opracownej metody z wynikami generowanymi przez algorytm GradCAM, wykazując wyższość opracowanego algorytmu w określaniu szczegółowych cech, na które zwraca uwagę sieć neuronowa podczas klasyfikacji. Ponadto zaprezentowano możliwość wykorzystania klasteryzacji do przedstawienia podobieństwa klas za pomocą wykresów 2D. W ramach kolejnego eksperymentu zweryfikowano, że opracowana metoda może być z powodzeniem stosowana do wszystkich klasycznych architektur splotowych sieci neuronowych. Zweryfikowano również, że metoda jest odporna na obroty obrazu wejściowego. Wszystkie przeprowadzone badania zaprezentowano w sposób przejrzysty oraz kompletny. Można mieć jednak pewną uwagę do doboru metod, z którymi porównywano zaproponowane algorytmy ponieważ zabrakło takich algorytmów jak GradCAM++, Guided GradCAM oraz Gradient Backpropagation

Rozdział 5 poświęcono na podsumowanie pracy i prezentację wniosków.

3. Poprawność i oryginalność postawionych zagadnień i stopień w jakim one zostały rozwiązane

W ramach pracy rozwiązano 4 oryginalne zagadnienia naukowe, które dotyczyły odpowiednio:

- opracowania metody wyznaczania map atencji dla splotowych sieci neuronowych, zapewniającej większą rozdzielczość w wizualizacji istotnych cech niż klasyczne algorytmy,
- opracowania metody do oceny i interpretacji wyników uzyskanych za pomocą map atencji z wykorzystaniem miary ilościowej,
- opracowanie kompleksowej metody do oceny jakości metod wyznaczających mapy atencji dla splotowych sieci neuronowych,
- opracowania metody do oznaczenia na obrazach wejściowych cech, które generują w warstwach splotowych podobne wzorce aktywności.

Wszystkie postawione powyżej zagadnienia naukowe uznaję za aktualne, poprawne oraz oryginalne. Użyte metody oraz przyjęte założenia uważam za poprawne i uzasadnione. Doktorant wykazał się dużą znajomością stanu wiedzy i literatury w zakresie metod wizualizacji atencji splotowych sieci neuronowych co pozwoliło mu celnie zidentyfikować aktualne problemy tych metod i w efekcie zaproponować własne rozwiązania przesuwające stan wiedzy w tym obszarze naukowym o kolejny krok do przodu. Należy zwrócić uwagę, że podjęte zagadnienia są bardzo istotne dla dalszego rozwoju głębokich sieci neuronowych. Pomimo ogromnych sukcesów uczenia głębokiego w zagadnieniach rozpoznawania obrazów i przetwarzania języka naturalnego pojawiają się również krytyczne raporty wskazujące na problemy z wyjaśnieniem decyzji podejmowanych przez głębokie sieci neuronowe w szczególności w zastosowaniach do diagnostyki medycznej. Ponadto istotnym problemem głębokich sieci neuronowych jest bardzo trudna analiza przyczyn błędnych decyzji. Metody wypracowane przez Doktoranta pomagają po pierwsze zidentyfikować problemy, które nie są widoczne przy klasycznej walidacji oraz zrozumieć przyczynę tych problemów co w efekcie pozwoli im przeciwdziałać.
W toku przeprowadzonych prac badawczych Doktorant rozwiązał w pełni wszystkie postawione zagadnienia. Wyniki zaprezentowanych doświadczeń wskazują, że zaproponowane metody posiadają pożądane właściwości oraz charakteryzują się lepszymi wynikami niż metody znane z literatury. Uzyskane wyniki zostały szczegółowo omówione z wykorzystaniem dużej ilości ilustracji. Należy nadmienić, że opracowane metody zostały zaimplementowane i udostępnione do użytku publicznego.

4. Analiza źródeł

Bibliografia zawiera 108 pozycji literaturowych. W większości są to artykuły z recenzowanych czasopism o zasięgu międzynarodowym oraz uznanych międzynarodowych konferencji. Pozycje literaturowe są aktualne i dobrze dobrane. Potwierdza to, że doktorant spędził dużo czasu na analizie bieżącego stanu wiedzy w zakresie zagadnień wokół, których prowadził badania. W ramach rozprawy doktorant posługiwał się głównie źródłami literaturowymi z obszaru uczenia głębokiego, uczenia maszynowego oraz wizualizacji atencji głębokich sieci neuronowych.

5. Pozycja rozprawy w stosunku do stanu wiedzy i stanu techniki reprezentowanych przez literaturę światową oraz znaczenie uzyskanych wyników dla dyscypliny naukowej

Praca Doktoranta wpisuje się w bardzo aktywny obecnie obszar prac badawczych w zakresie wytłumaczalnej (interpretowalnej) sztucznej inteligencji (ang. Explainable Artificial Intelligence, XAI). Prace w tym zakresie są między innymi niezmiernie ważne dla dalszego rozwoju technik uczenia głębokiego. Zaproponowane w rozprawie metody do wizualizacji i oceny atencji splotowych sieci neuronowych są krokiem w przód w stosunku

do tego co do tej pory zaproponowano w tym obszarze w literaturze światowej. Praca wnosi zatem istotny wkład w rozwój metod uczenia głębokiego, które w dyscyplinie Informatyka Techniczna i Telekomunikacja zajmują ważne miejsce.
W szczególności za duże osiągnięcie należy uznać opracowanie metody oceny ilościowej jakości map atencji dla splotowych sieci neuronowych. Zaproponowane rozwiązanie jest nowatorskie ponieważ pozwala na zautomatyzowane wykrywanie problemów z klasyfikacją w rzeczywistych problemach gdzie mamy do czynienia z bardzo dużą ilością klas oraz obrazów. Dotychczasowe metody wymagały wizualnej oceny uzyskanych wyników przez co ich użyteczność ogranicza się do walidacji tylko niewielkich zbiorów danych.
Za równie istotne osiągnięcie na tle literatury światowej uznaję metodę do oznaczenia na obrazie wejściowym wartości aktywacji wysokopoziomowych cech z warstw splotowych (metoda PRISM). Pozwala ona na zaawansowaną analizę i interpretację wyników uzyskiwanych za pomocą splotowych sieci neuronowych do tej pory niedostępną.
Autorowi udało się również ulepszyć metodę odkrywania szczegółowych cech obiektów, na których skupiają swoją uwagę splotowe sieci neuronowe podczas klasyfikacji. Eksperymentalnie wykazał, że uzyskane wyniki są lepsze niż dla uznanych metod z literatury światowej.
O wysokim, światowym poziomie prowadzonych prac badawczych przez Doktoranta świadczą również jego publikacje w czasopiśmie IEEE Access z listy filadelfijskiej z IF=3,367, w 4 monografiach wydawnictwa Springer oraz na międzynarodowej konferencji International Conference on Computational Science w 2022 roku.

6. **Umiejętność autora do poprawnego i przekonywującego przedstawienia uzyskanych wyników**

Główne osiągnięcia Doktoranta zostały zaprezentowane w rozdziale 3 oraz 4. O ile do rozdziału 4 nie mam większych zastrzeżeń ponieważ wyniki prac doświadczalnych zostały generalnie dobrze zaprezentowane to do części teoretycznej w rozdziale 3 mam uwagę dotyczącą zbyt zdawkowych opisów zaproponowanych metod. Wprawdzie są one wystarczające do zrozumienia koncepcji proponowanych rozwiązań jednak tekst techniczny powinien być precyzyjny i prezentować wszelkie szczegóły proponowanych rozwiązań. W ramach pracy przeprowadzono wiele prac doświadczalnych, które miały na celu wykazania poprawności opracowanych metod. Jako istotną zaletę pracy należy uznać fakt, że Autor nie waha się formułować krytycznych uwag odnośnie swoich rozwiązań prezentując ich ograniczenia i wady.
Biorąc pod uwagę całokształt, pomimo pewnych mankamentów pracy, które szczegółowo sformułowałem w punkcie 7 i 8, uznaję, że Autor w sposób przekonujący i poprawny przedstawił uzyskane wyniki.

7. **Główne wady i słabe strony rozprawy**

Rozprawę doktorską jako całość oceniam pozytywnie jednak jak prawie każda praca również ta posiada pewne niedociągnięcia lub uchybienia. Świadomość słabszych stron pracy pozwoli uniknąć podobnych błędów w przyszłości. Uwagi dotyczą organizacji niektórych rozdziałów oraz problemów do dalszej dyskusji naukowej.

Za słabą stronę rozprawy uznaję sposób prezentacji zaproponowanych metod w rozdziale 3. Według mojej oceny prezentacja tych metod jest zbyt pobieżna. Wprawdzie z przedstawionego opisu bez problemu można wywnioskować koncepcję poszczególnych

metod to jednak szczegółowe odtworzenie i implementacja metod byłoby już problemem. Zamieszczenie publicznie dostępnego kodu z implementacjami metod rekompensuje wspomnianą wadę tylko częściowo. Praca zyskałaby gdyby zamieszczono schematy, przykłady ilustrujące działanie opracowanych metod oraz wnikliwe opisano wszystkie istotne kroki poszczególnych metod co jest szczególnie wymagane w pracach z obszaru nauk technicznych.

Kolejną słabość pracy upatruję w niewielkiej ilości przykładów ilustrujących uzyskane wyniki w rozdziale 4. Biorąc pod uwagę, że praca ma 91 stron to nie byłoby problemem gdyby zamieszczono np. w formie dodatku większą ilość obrazów ilustrujących właściwości oraz problemy zaproponowanych podejść.

Istotną wadą rozprawy jest brak spisu rysunków i tabel, których w całej pracy jest stosunkowo dużo i w efekcie przemieszczanie się pomiędzy nimi jest utrudnione.

Praca posiada stosunkowo dużą liczbę poważnych błędów edytorskich, które zostały szczegółowo wymienione w punkcie 8.

Kolejna uwaga dotyczy algorytmu klasteryzacji cech uzyskanych za pomocą algorytmu PRISM przy użyciu Samoorganizującej się Mapy Cech (ang. Self-Organizing Map, SOM). Wspomniany mechanizm został opisany w podrozdziale 1.3.5 oraz 4.2.2.2. Niestety zrozumienie tego podejścia na podstawie zamieszczonych opisów nastręcza problemy. Brakuje informacji jak zdefiniowano wejście dla SOM oraz jaką przyjęto topologię i rozmiar sieci. Ponadto należało zaprezentować parametry uczenia sieci mimo, że nie jest to kluczowa sprawa z punktu widzenia wykorzystania tego algorytmu do klasteryzacji. Doprecyzowania również wymagają pojęcia "inclusive" oraz "exclusive" użyte w kontekście opisu cech.

Wyjaśnienia wymaga opis na str. 43 do wzoru 3. Autor pisze, że "*These gradients flowing back are global-average-pooled to obtain the importance weights for pixel k with respect to class c:*". Jednak indeks k oznacza numer kolejnej mapy cech z aktywacjami w warstwie splotowej, nie jest to indeks piksela z tej warstwy ponieważ za to są odpowiedzialne indeksy (i, j). Podobnie w przypadku zdania "*The importance weights create a saliency map as a matrix of values between 0.0 to 1.0 which corresponds to the importance of a particular pixel.*", które sugeruje, że mapa atencji powstaje bezpośrednio z wag wyznaczonych za pomocą wzoru 3. Niniejsze wagi są wykorzystywane tylko do tego aby "złożyć" za pomocą liniowej kombinacji map aktywacji wynikową mapę atencji. W tym miejscu pracy brakuje wzoru 2 z publikacji *(Selvaraju et al., 2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.*

Wzór 4 na str. 43 został nieprawidłowo zdefiniowany. Sumowanie w tym wzorze powinno odbywać się po elementach mapy atencji wyznaczonej wzorem 2 z pracy *Selvaraju et al., 2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization* a nie po α^c_k. Ponadto przy symbolu sumy pojawiają się indeksy x,y, których w ogóle nie ma przy sumowanej zmiennej.

Doprecyzowania wymaga jakiego typu detektor obiektów został użyty z systemu Detectron2 do implementacji algorytmu *Attention Focus Evaluation* oraz *Latent Features Detection*.

Pogłębionej dyskusji wymaga wpływ tego, że referencyjne obiekty są wyznaczane przez Detectron2 jako prostopadłe otoczki tzw. bounding box-y a obiekty wyznaczane za pomocą *Attention Focus Evaluation* mają w przybliżeniu realne kształty wykrytego obiektu. W związku z tym dopasowanie nawet w idealnym przypadku nie przyjmie wartości 100%.

W podrozdziale 3.2 zaprezentowano ideę algorytmu *Gradual Extrapolation*. Jednak opisano tylko czystą mechanikę działania tego algorytmu, zabrakło przedstawienia ogólnej koncepcji jaka przyświecała Autorowi algorytmu. Pożądane jest ogólne wyjaśnienie dlaczego zaproponowane podejście miałoby być lepsze niż istniejące metody.

Ponadto dla metody *Gradual Extrapolation* wymagane jest doprecyzowanie jak algorytm wstecznego propagowania mapy atencji przetwarza mapę gdy napotka warstwę splotową. W pracy wyjaśniono tylko zachowanie algorytmu dla warstwy max pooling.

W eksperymentach porównawczych prowadzonych w rozdziale 4 nie wykorzystano metod GradCAM++ oraz Guided GradCAM.

Powyższe uwagi nie podważają pozytywnej oceny rozprawy i jej znaczącego wkładu w rozwój dyscypliny naukowej.

8. Krytyczne uwagi szczegółowe na temat błędów o charakterze edytorskim

Dużym ułatwieniem dla czytelnika byłoby gdyby przygotowano spis ilustracji i tabel, których w pracy jest stosunkowo dużo.

streszczenie w języku polskim: W zdaniu "W pierwszej metodzie wprowadzam liczbę ocenę skupienia sieci na obiekcie" chodziło chyba o "liczbową ocenę" lub "ilościową ocenę".

str. 13: Zdanie "The network may learn only a small trait if an object instead it as a whole." jest mało zrozumiałe.

str. 14: Niezrozumiały fragment zdania "... plays a joyr role.".

str. 14: Brak odwołania w tekście do rysunku 5.

str. 6: Na rysunku 7 nie wyjaśniono co oznaczają zmienne x i y na osiach wykresu oraz jakie znaczenie mają kolory wykresu.

str. 18: Pojawia się słowo "meth" a powinno być "method".

str. 18: Brak litery w słowie "evn", powinno być chyba "even".

str. 19: Brak odwołania w tekście do rysunku 9.

str. 19: Jakość zapożyczonego rysunku 9 jest niezadowalająca, rysunek jest zbyt rozmyty.

str 20: Odwołanie do nieprawidłowego rysunku w zdaniu "Figure 3 illustrates that the stack of mathematical operations in the convolution layer completes the linear operation."

str. 20: Jakość zapożyczonego rysunku 10 jest niezadowalająca, rysunek jest zbyt rozmyty.

str. 25: Brak odwołania w tekście do rysunku 11.

str. 26: Brak odwołania w tekście do rysunku 12.

str. 27: Brakuje informacji w opisie rysunku 13, że został on zapożyczony z pracy (Zeiler and Fergus, 2014).

str. 27: Brak odwołania w tekście do rysunku 13.

str. 28: Brak odwołania w tekście do rysunku 14.

str. 29: Brak odwołania w tekście do rysunku 15.

str. 30: Brak odwołania w tekście do rysunku 16.

str. 32: Brak odwołania w tekście do rysunku 17.

str. 33: Brak odwołania w tekście do rysunku 18.

str. 35: Brak odwołania w tekście do rysunku 19.

str. 36: Brak odwołania w tekście do rysunku 20.

str. 42: Odwołanie do nieprawidłowego rysunku w zdaniu "Detectron2 produces images with highlighted ROIs (see fig. 1) and returns ...".

str 42: Niespójność nazewnictwa metody GradCam, pojawia sie nazwa "GradCam" oraz "Grad-CAM".

str. 47: Odwołanie do nieprawidłowego rysunku w zdaniu "... and truncates the results after the third Principal Component thus receiving an RGB map of features as seen in picture 1.".

str. 47: Brak odwołania w tekście do rysunku 25.

str. 48: Odwołanie do nieprawidłowego rysunku w zdaniu "This procedure results in a checkered representation of the processed images (fig. 3.)."

str. 56: Brak odwołania w tekście do rysunku 31.

str. 56: W opisie rysunku 31 i 32 nie zamieszczono informacji jakie wyniki przedstawiają poszczególne obrazy.

str. 57: Brak odwołania w tekście do rysunku 33.

str. 58: Brak odwołania w tekście do rysunku 35.

str. 59: Brak odwołania w tekście do rysunku 36.

str. 60: W tabeli w sekcji "Procedure" numeracja punktów zaczyna się od 8 zamiast od 1.

str. 61: Nieprawidłowa referencja do rysunku 37 w zdaniu "In Figure 6, Gradual Extrapolation offers noticeable improvements over conventional Grad-CAM and a minor improvement over similar Excitation Backpropagation.".

str. 62: Brak odwołania w tekście do tabeli 5.

str. 64: Brak odwołania w tekście do rysunku 38.

str. 67: W tabeli w sekcji "Procedure" numeracja punktów zaczyna sie od 8 zamiast od 1.

str. 68: Pojawia się słowo "Pawns" a powinno być "Paws".

str. 70: Zdanie "Despite that PRISM shines while processing several images at once, mostly in conjunction with GE, we believe it may also work as an indicator for ambivalent classes if combined with clustering technique, e.g. Self-Organizing Maps (SOM)." jest dla mnie niezbyt zrozumiałe.

str. 71: Na rysunku 44 nie opisano osi.

str. 72: Brak odwołania w tekście do tabeli 6.

str. 73: Brak odwołania w tekście do tabeli 7.

str. 76: Odwołanie w tekście do rysunku 45 bez podania jego numeru.

str. 82: Brak odwołania w tekście do rysunku 48.

9. Wnioski końcowe

Recenzowana rozprawa doktorska podejmuje oryginalny oraz istotny z punktu widzenia dyscypliny Informatyka Techniczna i Telekomunikacja problem badawczy. Wyniki przedstawione w rozprawie potwierdzają zagadnienia badawcze podjęte w ramach pracy i wnoszą oryginalny wkład do dyscypliny. Doktorant udowodnił, że potrafi rozwiązywać nietrywialne problemy naukowe stosując przyjęte w nauce metody poznawcze i badawcze.

Podsumowując stwierdzam, że recenzowana rozprawa doktorska spełnia wszystkie wymogi stawiane rozprawom doktorskim przez obowiązujące przepisy. W związku z powyższym wnioskuję o dopuszczenie recenzowanej rozprawy doktorskiej do publicznej obrony.

Marek Kowal