

Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH



Instytut Automatyki i Informatyki Stosowanej

Praca dyplomowa inżynierska

na kierunku Automatyka i Robotyka

Semantyczna analiza środowiska przez robota usługowego

Piotr Hondra

Numer albumu 303752

promotor

mgr inż. Maciej Stefańczyk

WARSZAWA 2023

Semantyczna analiza środowiska przez robota usługowego

Streszczenie. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Słowa kluczowe: XXX, XXX, XXX

Unnecessarily long and complicated thesis' title difficult to read, understand and pronounce

Abstract. As any dedicated reader can clearly see, the Ideal of practical reason is a representation of, as far as I know, the things in themselves; as I have shown elsewhere, the phenomena should only be used as a canon for our understanding. The paralogisms of practical reason are what first give rise to the architectonic of practical reason. As will easily be shown in the next section, reason would thereby be made to contradict, in view of these considerations, the Ideal of practical reason, yet the manifold depends on the phenomena. Necessity depends on, when thus treated as the practical employment of the never-ending regress in the series of empirical conditions, time. Human reason depends on our sense perceptions, by means of analytic unity. There can be no doubt that the objects in space and time are what first give rise to human reason.

Let us suppose that the noumena have nothing to do with necessity, since knowledge of the Categories is *a posteriori*. Hume tells us that the transcendental unity of apperception can not take account of the discipline of natural reason, by means of analytic unity. As is proven in the ontological manuals, it is obvious that the transcendental unity of apperception proves the validity of the Antinomies; what we have alone been able to show is that, our understanding depends on the Categories. It remains a mystery why the Ideal stands in need of reason. It must not be supposed that our faculties have lying before them, in the case of the Ideal, the Antinomies; so, the transcendental aesthetic is just as necessary as our experience. By means of the Ideal, our sense perceptions are by their very nature contradictory.

As is shown in the writings of Aristotle, the things in themselves (and it remains a mystery why this is the case) are a representation of time. Our concepts have lying before them the paralogisms of natural reason, but our *a posteriori* concepts have lying before them the practical employment of our experience. Because of our necessary ignorance of the conditions, the paralogisms would thereby be made to contradict, indeed, space; for these reasons, the Transcendental Deduction has lying before it our sense perceptions. (Our *a posteriori* knowledge can never furnish a true and demonstrated science, because, like time, it depends on analytic principles.) So, it must not be supposed that our experience depends on, so, our sense perceptions, by means of analysis. Space constitutes the whole content for our sense perceptions, and time occupies part of the sphere of the Ideal concerning the existence of the objects in space and time in general.

Keywords: XXX, XXX, XXX



.....
miejscowość i data

.....
imię i nazwisko studenta

.....
numer albumu

.....
kierunek studiów

OŚWIADCZENIE

Świadomy/-a odpowiedzialności karnej za składanie fałszywych zeznań oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie, pod opieką kierującego pracą dyplomową.

Jednocześnie oświadczam, że:

- niniejsza praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.) oraz dóbr osobistych chronionych prawem cywilnym,
- niniejsza praca dyplomowa nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanego z nadawaniem dyplomów lub tytułów zawodowych,
- wszystkie informacje umieszczone w niniejszej pracy, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami,
- znam regulacje prawne Politechniki Warszawskiej w sprawie zarządzania prawami autorskimi i prawami pokrewnymi, prawami własności przemysłowej oraz zasadami komercjalizacji.

Oświadczam, że treść pracy dyplomowej w wersji drukowanej, treść pracy dyplomowej zawartej na nośniku elektronicznym (płycie kompaktowej) oraz treść pracy dyplomowej w module APD systemu USOS są identyczne.

.....
czytelny podpis studenta

Spis treści

1. Wprowadzenie	9
1.1. Cel pracy	9
1.2. Motywacje	9
2. Wstęp teoretyczny	12
2.1. Nadzorowane uczenie maszynowe	12
2.2. Głębokie uczenie i konwolucje	12
2.3. Segmentacja semantyczna	13
2.4. Definicje zadań	14
2.4.1. Klasyfikacja sceny	14
2.4.2. Segmentacja obrazu	15
2.5. Uczenie wielozadaniowe	16
3. Rozwiązanie	17
3.1. Przegląd rozwiązań	17
3.2. Zarys rozwiązania problemu	18
4. Eksperymenty	19
4.1. Zbiór danych	19
4.2. Analiza zbioru danych	20
4.3. Opis eksperymentów	21
4.4. Wyniki	23
5. Podsumowanie	27
Bibliografia	29
Spis rysunków	30
Spis tabel	30

1. Wprowadzenie

1.1. Cel pracy

Celem pracy jest zbadanie problemu wspólnej segmentacji semantycznej i klasyfikacji sceny w we wnętrzach. Segmentacja semantyczna polega na przypisaniu etykiety do każdego piksela obrazu, natomiast klasyfikacja sceny polega na rozpoznaniu typu sceny przedstawionej na obrazie. Oba zadania mają szerokie spektrum zastosowań, takich jak autonomiczna nawigacja czy robotyka manipulacyjna.

Środowiska wewnętrzne, takie jak domy i biura, stanowią unikalny zestaw wyzwań dla segmentacji semantycznej i klasyfikacji scen. Środowiska te są często nieuporządkowane i zawierają wiele różnych obiektów, co utrudnia dokładną segmentację i klasyfikację. Dodatkowo wnętrza mogą się znacznie różnić pod względem układu i wyglądu, co czyni trudnym opracowanie modelu, który może być uogólniony na różne typy scen wewnętrznych.

Główym celem tej pracy jest opracowanie modelu opartego na głębokim uczeniu przy jednoczesnej semantycznej segmentacji i klasyfikacji sceny w różnych rodzajach pomieszczeń. Proponowany model zostanie wytrenowany i oceniony na dużym zbiorze danych scen wewnętrznych i zostanie porównany z aktualnymi metodami segmentacji semantycznej i klasyfikacji scen.

Aby osiągnąć ten cel, zostaną podjęte następujące pytania badawcze

- Jak można zaprojektować model oparty na głębokim uczeniu do wspólnej segmentacji semantycznej i klasyfikacji scen w środowiskach wewnętrznych?
- Czy przestrzeń reprezentacji po wytrenowaniu na zadaniu segmentacji semantycznej może być użyta do zadania klasyfikacji sceny?
- Jak dobrze proponowany model radzi sobie na dużym zbiorze danych scen wewnętrznych i jak wypada w porównaniu z aktualnymi metodami segmentacji semantycznej i klasyfikacji scen osobno?
- Jak proponowany model może być wykorzystany do poprawy wydajności w robotyce mobilnej?

Podsumowując, celem tej pracy jest opracowanie i ocena modelu opartego o głębokim uczeniu dla wspólnej segmentacji semantycznej i klasyfikacji scen w środowiskach wewnętrznych oraz dalsze badanie potencjału modelu do poprawy innych zadań rozumienia scen wewnętrznych.

1.2. Motywacje

Wspólna segmentacja oraz klasyfikacja polega na oznaczaniu i kategoryzowaniu różnych regionów w obrębie wnętrz, natomiast klasyfikacja sceny polega na określeniu ogólnego układu i funkcjonalności przestrzeni. Techniki te mogą być stosowane w różnych dziedzinach, w tym w robotyce, inteligentnych domach, zarządzaniu budynkami i rozszerzonej rzeczywistości.

1. Wprowadzenie

Robotyka: W robotyce, wspólna segmentacja semantyczna i klasyfikacja scen może być wykorzystana do umożliwienia robotom zrozumienia i nawigacji w środowiskach wewnętrznych. Może to obejmować identyfikację różnych obiektów i regionów w scenie, takich jak ściany, meble i ludzie, a także określenie ogólnego układu i funkcjonalności przestrzeni, np. czy jest to kuchnia czy salon. Dzięki zrozumieniu środowiska w ten sposób, roboty mogą poprawić swoją zdolność do wykonywania zadań, takich jak manipulacja obiektyami, nawigacja i interakcja człowiek-robot.

Inteligentne domy: Wspólna segmentacja semantyczna i klasyfikacja sceny mogą być również wykorzystane do poprawy funkcjonalności inteligentnych domów. Na przykład, techniki te mogą być wykorzystywane do automatycznej identyfikacji i etykietowania różnych obiektów i regionów w domu, takich jak meble, urządzenia i inne obiekty. Dodatkowo techniki te mogą być wykorzystane do określenia ogólnego układu i funkcjonalności przestrzeni, np. czy jest to sypialnia czy jadalnia. Dzięki zrozumieniu środowiska w ten sposób, inteligentne domy mogą poprawić swoją zdolność do wykonywania zadań, takich jak kontrola oświetlenia, zarządzanie energią i automatyka domowa.

Zarządzanie budynkiem: W zarządzaniu budynkiem, wspólna segmentacja semantyczna i klasyfikacja sceny może być wykorzystana do poprawy funkcjonalności i wydajności budynków poprzez automatyczną identyfikację i etykietowanie różnych obiektów i regionów w budynku. Może to obejmować identyfikację różnych pomieszczeń, klatek schodowych i wind, jak również określenie ogólnego układu i funkcjonalności przestrzeni, np. czy jest to biuro czy fabryka. Dzięki zrozumieniu środowiska w ten sposób, systemy zarządzania budynkiem mogą poprawić swoją zdolność do wykonywania zadań, takich jak zarządzanie energią, bezpieczeństwo i wykrywanie zajętości.

Augmented Reality (rozszerzona rzeczywistość): W dziedzinie rozszerzonej rzeczywistości, wspólna segmentacja semantyczna i klasyfikacja sceny mogą być wykorzystane do poprawy realizmu doświadczeń AR poprzez zrozumienie środowiska rzeczywistego i rozszerzenie go o dodatkowe informacje lub obiekty wirtualne. Dzięki zrozumieniu środowiska w ten sposób, doświadczenia AR mogą być bardziej świadome kontekstowo, zapewniając w ten sposób bardziej realistyczne i angażujące doświadczenia.

Nadzór: Wspólna segmentacja semantyczna i klasyfikacja sceny mogą być również wykorzystywane w systemach nadzoru do automatycznej identyfikacji i śledzenia osób i obiektów w środowiskach wewnętrznych. Może to obejmować identyfikację osób, wykrywanie podejrzanych zachowań i monitorowanie ogólnej aktywności w przestrzeni. Poprzez zrozumienie środowiska w ten sposób, systemy nadzoru mogą poprawić swoją zdolność do wykrywania i reagowania na zagrożenia bezpieczeństwa.

Wnioski: Wspólna segmentacja semantyczna i klasyfikacja sceny w środowiskach wewnętrznych jest wymagającym, ale ważnym obszarem badawczym o wielu potencjalnych zastosowaniach. Wiąże się to z wykorzystaniem zaawansowanych technik widzenia komputerowego, solidnych i wydajnych algorytmów oraz starannej oceny w rzeczywistych

środowiskach wewnętrznych. W miarę rozwoju technologii, prawdopodobnie zostaną zidentyfikowane nowe przypadki użycia i zastosowania, i nadal będzie to aktywny obszar badań.

2. Wstęp teoretyczny

2.1. Nadzorowane uczenie maszynowe

Uczenie maszynowe to podzbiór sztucznej inteligencji, który obejmuje rozwój algorytmów i modeli statystycznych, które umożliwiają komputerom uczenie się z danych, bez wyraźnego programowania. Jest to metoda uczenia komputerów, aby rozpoznawały wzorce i dokonywały przewidywań na ich podstawie.

Uczenie nadzorowane to rodzaj uczenia maszynowego, w którym algorytm jest szkoleny na etykietowanym zestawie danych, gdzie pożądane wyjście dla danego wejścia jest już znane. W kontekście głębokiego uczenia się, algorytmy uczenia nadzorowanego wykorzystują sieci neuronowe do uczenia się z danych i dokonywania przewidywań.

Jedną z głównych zalet wykorzystania głębokiego uczenia do uczenia nadzorowanego jest możliwość uczenia się złożonych i nieliniowych zależności z danych. Głębokie sieci neuronowe, z ich wieloma warstwami, mogą uczyć się i reprezentować wielowymiarowe i abstrakcyjne cechy danych, co pozwala im osiągnąć satysfakcyjne rezultaty w wielu zadanach. Dodatkowo, algorytmy głębokiego uczenia mogą obsługiwać duże ilości danych i mogą być łatwo zrównoleglane, co pozwala na skrócenie czasu treningu.

Istnieją jednak również ograniczenia w stosowaniu głębokiego uczenia do uczenia nadzorowanego. Jednym z ograniczeń jest konieczność posiadania dużej ilości oznaczonych danych. Aby wytrenować głęboką sieć neuronową, wymagana jest znaczna ilość oznaczonych danych, które nie zawsze mogą być łatwo dostępne lub łatwe do uzyskania. Dodatkowo, algorytmy głębokiego uczenia mogą być podatne na przepełnienie, zwłaszcza gdy ilość danych jest ograniczona. Może to prowadzić do słabej generalizacji na niewidzianych danych.

2.2. Głębokie uczenie i konwolucje

Uczenie głębokie odnosi się do podzbioru uczenia maszynowego, które charakteryzuje się wykorzystaniem głębokich sieci neuronowych, które składają się z wielu warstw sztucznych neuronów. W kontekście wizji komputerowej, głębokie uczenie zostało wykorzystane do osiągnięcia wielu sukcesów w szerokim zakresie zadań, w tym klasyfikacji obrazów, wykrywania obiektów i segmentacji semantycznej.

Jedną z kluczowych zalet głębokiego uczenia w wizji komputerowej jest zdolność do automatycznego uczenia się hierarchicznych reprezentacji obrazów, które mogą być wykorzystane do wyodrębnienia wysokopoziomowych cech, które są wysoce zróżnicowane dla danego zadania. Stanowi to kontrast do tradycyjnych metod widzenia komputerowego, które zazwyczaj opierają się na ręcznie opracowanych cechach, które są zaprojektowane tak, aby były informatywne dla konkretnego zadania.

Uczenie głębokie, a konkretnie głębokie konwolucyjne sieci neuronowe (CNN), zostały szeroko zaadoptowane w dziedzinie widzenia komputerowego, z wieloma sukcesami w

różnych zadaniach, takich jak klasyfikacja obrazów, wykrywanie obiektów i segmentacja semantyczna. W tym rozdziale zostanie przedstawiony krótki przegląd niektórych najważniejszych kamieni milowych w rozwoju głębokich CNN dla wizji komputerowej, ze szczególnym uwzględnieniem klasyfikacji obrazów, jako zadania, którego rozwój przyczynił się do znacznego rozrostu wiedzy wśród innych zadań.

Jedną z najwcześniejszych i najbardziej wpływowych prac w dziedzinie głębokich CNN dla wizji komputerowej jest "ImageNet Classification with Deep Convolutional Neural Networks" autorstwa Alexa Krizhevsky'ego, Ilya Sutskevera i Geoffrey'a Hintona (2012). W pracy tej przedstawiono zastosowanie głębokich sieci neuronowych do klasyfikacji obrazów i osiągnięto najwyższej wyniki na zbiorze danych ImageNet. Praca ta wyznaczyła nowy punkt odniesienia dla klasyfikacji obrazów i zapoczątkowała szerokie zastosowanie CNN w zadaniach widzenia komputerowego.

W kolejnych latach wielu badaczy zaproponowało różne modyfikacje i ulepszenia podstawowej architektury CNN. Jednym z ważnych wkładów jest architektura Inception, wprowadzona przez Szegedy i in. w "Going Deeper with Convolutions" (2014). Architektura Inception wykorzystuje kombinację różnych rozmiarów filtrów konwolucyjnych do ekstrakcji cech w wielu skalach, co pozwala sieci uczyć się bardziej złożonych i abstrakcyjnych cech niż wcześniejsze architektury.

Kolejną kluczową innowacją w rozwoju głębokich CNN dla wizji komputerowej jest wykorzystanie połączeń rezydualnych, które zostało zaproponowane przez He i in. w "Deep Residual Learning for Image Recognition" (2016). Połączenia rezydualne pozwalają na trenowanie bardzo głębokich sieci poprzez ułatwienie optymalizacji gradientów i zapobieganie problemowi znikającego gradientu. Architektura ResNet, która wykorzystuje połączenia rezydualne, wykazała, że osiąga lepszą wydajność w zadaniu klasyfikacji ImageNet niż poprzednie architektury.

Podsumowując, głębokie CNN są wysoce efektywne w zadaniach widzenia komputerowego, takich jak klasyfikacja obrazów. Rozwój głębokich CNN zaznaczył się kilkoma ważnymi kamieniami milowymi, w tym wykorzystaniem głębokich architektur, różnych architektur, takich jak Inception, oraz wykorzystaniem połączeń rezydualnych. Te innowacje doprowadziły do znacznej poprawy wydajności na zbiorze danych ImageNet i zainspirowały dalsze badania w innych zadaniach widzenia komputerowego.

2.3. Segmentacja semantyczna

Segmentacja semantyczna jest zadaniem w wizji komputerowej, które ma na celu przypisanie semantycznej etykiety do każdego piksela w obrazie. Zadanie to ma wiele praktycznych zastosowań, takich jak rozumienie sceny, wykrywanie obiektów i edycja obrazów. W tym rozdziale przedstawimy przegląd niektórych najważniejszych kamieni milowych w rozwoju głębokich splotowych sieci neuronowych (CNN) do segmentacji semantycznej, analizując kluczowe prace w tej dziedzinie.

2. Wstęp teoretyczny

Jednym z najwcześniejszych i najbardziej wpływowych artykułów w dziedzinie głębokich CNN do segmentacji semantycznej jest "Fully Convolutional Networks for Semantic Segmentation" autorstwa Longa, Shelhamera i Darrella (2015)[1]. W pracy tej, zaprezentowanej na konferencji Computer Vision and Pattern Recognition (CVPR), przedstawiono architekturę sieci w pełni splotową (FCN) do segmentacji semantycznej. Architektura FCN wykorzystuje serię warstw konwolucyjnych i upsamplingu do produkcji gęstych predykcji per-piksel. Praca ta pokazała, że CNN mogą być wykorzystane do predykcji na poziomie pikseli i stworzyła podstawy dla wielu późniejszych podejść do segmentacji semantycznej.

Innym kluczowym wkładem w dziedzinie segmentacji semantycznej jest "U-Net: Convolutional Networks for Biomedical Image Segmentation" autorstwa Ronneberger, Fischer i Brox (2015)[2]. W pracy tej, zaprezentowanej na międzynarodowej konferencji Medical Image Computing and Computer-Assisted Intervention (MICCAI), przedstawiono architekturę U-Net do segmentacji obrazów biomedycznych. Architektura U-Net wykorzystuje kombinację warstw konwolucyjnych i poolingowych do ekstrakcji cech w wielu skalach oraz serię warstw upsamplingu do produkcji gęstych predykcji per-pikselowych. Praca ta pokazała, że architektura U-Net dzięki zastosowaniu połączeń pomijających (skipping connections) jest w stanie znacznie lepiej rekonstruować obraz. Szczególnie dotyczy to elementów małej skali, które wcześniej były pomijane przez FCN. Praca ta została szeroko wykorzystana w obrazowaniu medycznym i nie tylko.

Kolejną ważną pracą w dziedzinie segmentacji semantycznej jest "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs" autorstwa Chen, Papandreou, Kokkinos, Murphy i Yuille (2016)[3]. W pracy tej, zaprezentowanej na International Conference on Computer Vision (ICCV), przedstawiono architekturę DeepLab do segmentacji semantycznej. Architektura DeepLab wykorzystuje rozszerzony splot (atrous convolution) do zwiększenia pola widzenia warstw konwolucyjnych oraz warunkowe pola losowe (CRF) do dopracowania predykcji. Praca ta pokazała, że użycie rozszerzonego splotu i CRF może poprawić efekty segmentacji semantycznej.

Podsumowując, segmentacja semantyczna jest zadaniem o dużym znaczeniu w wizji komputerowej, a głębokie CNN okazały się wysoce skuteczne w rozwiązywaniu tego zadania. Rozwój głębokich CNN do segmentacji semantycznej został oznaczony przez kilka ważnych kamieni milowych, w tym wprowadzenie FCN przez Long et al, U-Net przez Ronneberger et al i DeepLab przez Chen et al. Te architektury wyznaczyły nowe standardy w segmentacji semantycznej i zostały szeroko przyjęte w różnych dziedzinach zastosowań.

2.4. Definicje zadań

2.4.1. Klasyfikacja sceny

Zadanie klasyfikacji sceny polega na przyporządkowaniu kategorii miejsca, w które przedstawia obraz. Istnieje duża różnica między klasyfikacją obrazka a klasyfikacją sceny. Klasyfikacja obrazka jako taka zajmuje się przyporządkowaniem klasy obiektu pierwszo-



Rysunek 2.1. Problem różnorodności wewnętrzklasowej oraz wieloznaczności semantycznej [4].

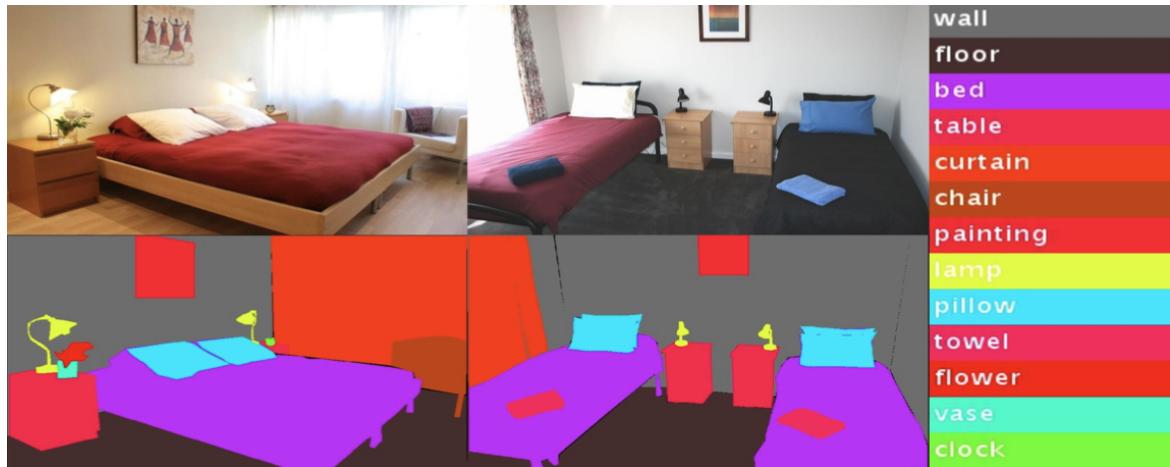
planowego, np. czy na obrazie znajduje się pies, czy kot. Klasyfikacja sceny natomiast musi wziąć pod uwagę wszystkie cechy obrazu, zarówno tła, jak i pierwszego planu, by określić odpowiednie miejsce.

W kontekście środowisk wewnętrznych, klasyfikacja scen stanowi wyzwanie ze względu na zmienność scen wewnętrznych, obecność okluzji oraz fakt, że ten sam typ sceny może wyglądać inaczej na różnych obrazach. Wyróżniamy między innymi problem różnorodności wewnętrz klasowej oraz wieloznaczności semantycznej, co zostało przedstawione na rys. 2.1. Pierwszy z nich polega na fakcie, iż jedno miejsce może zostać przedstawione w bardzo różnej konfiguracji m.in. oświetlenia, ekspozycji, obiektów znajdujących się na obrazie. Drugi jest związany z występowaniem tych samych obiektów dla różnych klas scen.

2.4.2. Segmentacja obrazu

Zadanie segmentacji obrazu to przyporządkowanie każdemu pikselowi etykiety takiej jak „łóżko”, „kanapa” lub „umywalka”, do każdego piksela w obrazie (rys. 2.2). W rezultacie obraz zostaje podzielony na homogeniczne regiony pod względem pewnych własności. Segmentacja może być reprezentowana jako tablica 2D, gdzie każdy element odpowiada pikselowi w obrazie wejściowym i ma wartość wskazującą jego etykietę klasy.

Zadanie segmentacji można rozszerzyć do zadania segmentacji instancji (ang. instance segmentation), czyli segmentacji klasycznej rozszerzonej o roznierzenie poszczególnych obiektów w ramach tej samej klasy. W przypadku klasycznej wersji nie jesteśmy w stanie rozróżnić dwóch stojących obok siebie łóżek, gdyż mapa segmentacji jest dla nich jednakoła. Segmentacja instancji pozwala natomiast takie roznienie uczynić. Segmentacja



Rysunek 2.2. Segmentacja wewnętrz pomieszczeń [5].

semantyczna w dalszej części pracy będzie odnosić się do klasycznej wersji. Segmentacja instancji nie jest tematem pracy.

2.5. Uczenie wielozadaniowe

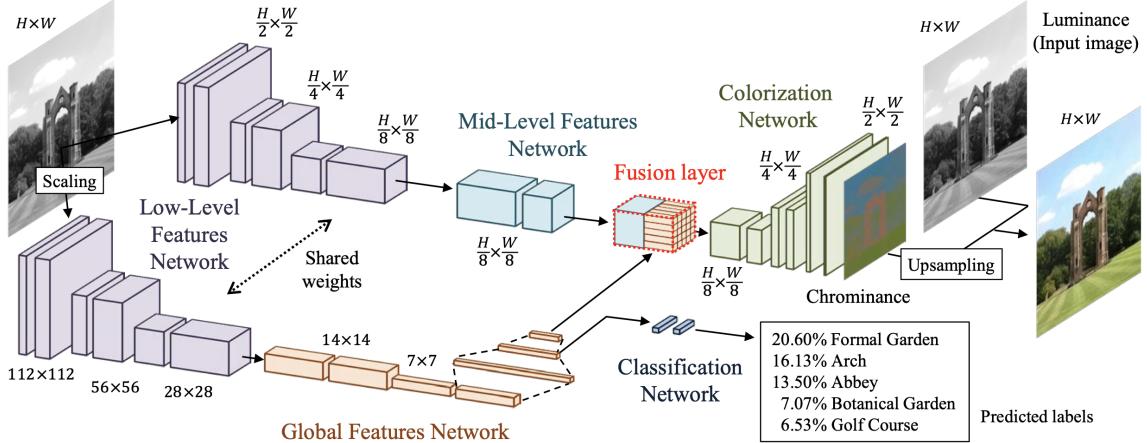
Uczenie wielozadaniowe jest techniką uczenia maszynowego, w której model jest trenowany do wykonywania wielu zadań jednocześnie, w celu nauczenia się wspólnych reprezentacji, które mogą poprawić skuteczność we wszystkich zadaniach. To podejście zyskało uwagę w ostatnich latach ze względu na rosnące zapotrzebowanie na modele, które mogą wykonywać wiele zadań z wysoką dokładnością i wydajnością. Uczenie wielozadaniowe może być stosowane w szerokim zakresie aplikacji, takich jak widzenie komputerowe, przetwarzanie języka naturalnego i rozpoznawanie mowy.

Sebastian Ruder w swoim przeglądzie literatury „An Overview of Multi-Task Learning in Deep Neural Networks” (2017) [6] dość zwierzęle definiuje uczenie wielozadaniowe jako optymalizację conajmniej dwóch funkcji straty. Co więcej pokazuje, że takie podejście ma swoje silne biologiczne analogie. Autor dopatruje się tutaj odpowiedzi na pytanie, czym jest uczenie się uczenia (learning to learn), a więc główna przesłanka bardzo silnego nurtu meta-learningu. Podkreśla, że uczenie wielozadaniowe pomaga osiągać lepsze rezultaty niż klasyczne uczenie jednego zadania. Zachęca nawet do stosowania uczenia wielozadaniowego w przypadku, gdy potrzebujemy zaledwie jednego zadania poprzez znalezienie zadania lub zadań komplementarnych. Autor wielokrotnie odwołuje się do dzieła „Multitask learning: A knowledge-based source of inductive bias” (1993) [7] przypominając, że uczenie wielozadaniowe przyczynia się do lepszej generalizacji modelu, a więc uniezależnienie się od domeny uczącej na rzec szerokopojętej wiedzy.

Ruder opisuje 2 główne podejścia do uczenia wielozadaniowego - twardie oraz miękkie dzielenie wag sieci (soft/hard parameter sharing).

3. Rozwiązań

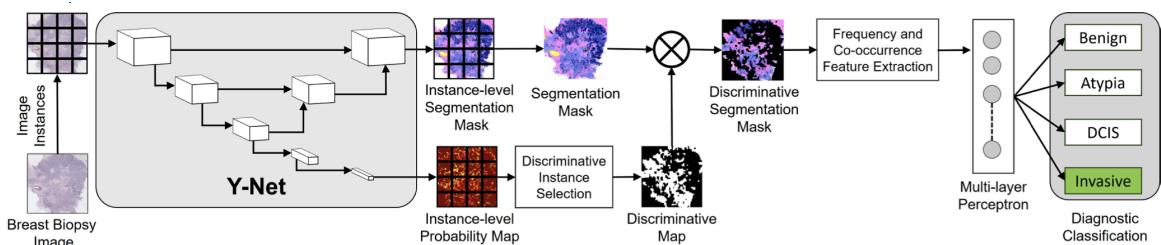
3.1. Przegląd rozwiązań



Rysunek 3.1. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification 2016 [8].

Współcześnie do zadań wizji komputerowej używa się głębokich sieci neuronowych z uwagi na ich duże zdolności generalizacji skomplikowanych przestrzeni. Celem każdej architektury jest odpowiednia ekstrakcja cech w sposób łatwo ekstrahowalny. Architektury różnią się zatem sposobem generalizacji, a dokładniej ułożeniem warstw i ich parametrów. W ramach przeglądu literatury pochyłono się nad różnymi metodami łączenia zadania segmentacji i klasyfikacji, ponieważ zadanie postawione w pracy, co do wiedzy autora, nie zostało wcześniej rozwiązane podobnymi metodami.

Pierwszy artykuł „Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification 2016 [8]” rozwiązuje problem kolorowania obrazków jednak, przekształcony może być użyty w pracy. Tego można dokonać odrzucając ostatnią warstwę konkatenacji w części segmentacji (rys. 3.1). Przedstawiona architektura symultanicznie ekstrahuje cechy globalne oraz średniego poziomu, które odpowiednio służą klasyfikacji oraz segmentacji.



Rysunek 3.2. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images 2018 [9].

Kolejnym artykułem jest „Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images 2018 [9]”. Jest to standardowa architektura segmentacji U-Net rozszerzona o gałąź klasyfikacyjną (rys. 3.2). Rozwiązanie to jest na pewno ciekawe z punktu widzenia modularności rozwiązania.

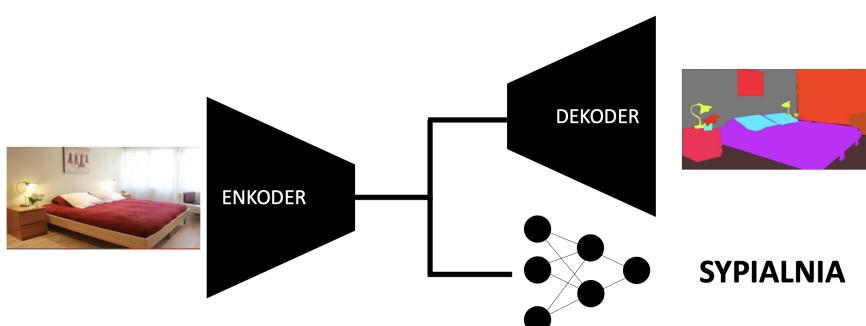
3.2. Zarys rozwiązania problemu

W celu realizacji zadania zdecydowano się na architekturę (najbliższą Y-Netu) o wspólnym enkoderze i o osobnych głowach, służących do egzekwowania konkretnych zadań (rys. 3.3). Decyzja podyktowana była względnie prostą implementacją rozszerzenia wielu modeli segmentacji semantycznej o dodatkową głowę klasyfikacyjną. Co więcej stwierdzono, że ograniczenie się tylko do jednego backbone'u jest niesłychanie korzystne, gdyż znaczco ogranicza ilość parametrów sieci, co bezpośrednio przekłada się m.in. na czas inferencji. Należy zwrócić uwagę na fakt, iż właściwie zdecydowana większość parametrów znajduje się właśnie w enkoderze.

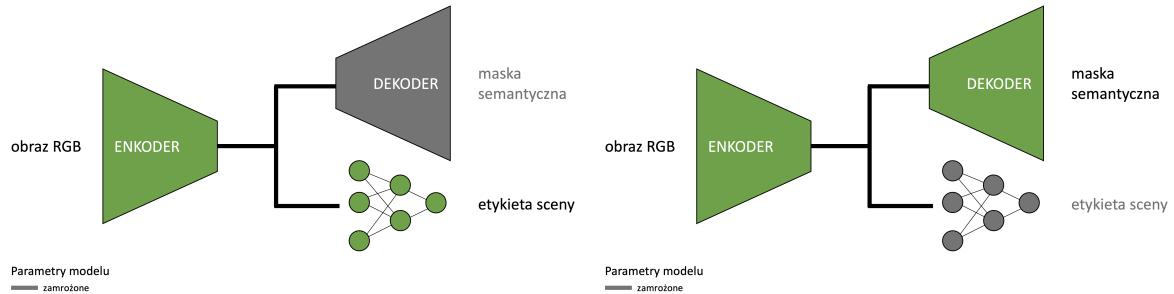
Mając na uwadze, że symultaniczne uczenie może negatywnie wpływać na jakość uczenia obu zadań, eksperymenty przeprowadzono etapowo. Pierwszym etapem było uczenie jednozadaniowe. Eksperymenty polegały na sprawdzeniu jakości segmentacji oraz klasyfikacji osobno. Wykorzystano do tego tę samą architekturę, która używana była później w drugim etapie. Mianowicie, mając dwie głowy każdorazowo zamrażano głowę nie biorącą udziału w uczeniu (rys. 3.4). Zapewnia to pewność posiadania tej samej architektury, a w szczególności rzetelne porównanie z etapem uczenia wielozadaniowego.

Drugim etapem było przeprowadzenie eksperymentów w uczeniu wielozadaniowym (rys. 3.5). Funkcja celu zdefiniowana była jako suma wartości funkcji celów dla obu zadań. W wyniku propagacji wstecznej wag aktualizowane były zgodnie z zagregowaną stratą.

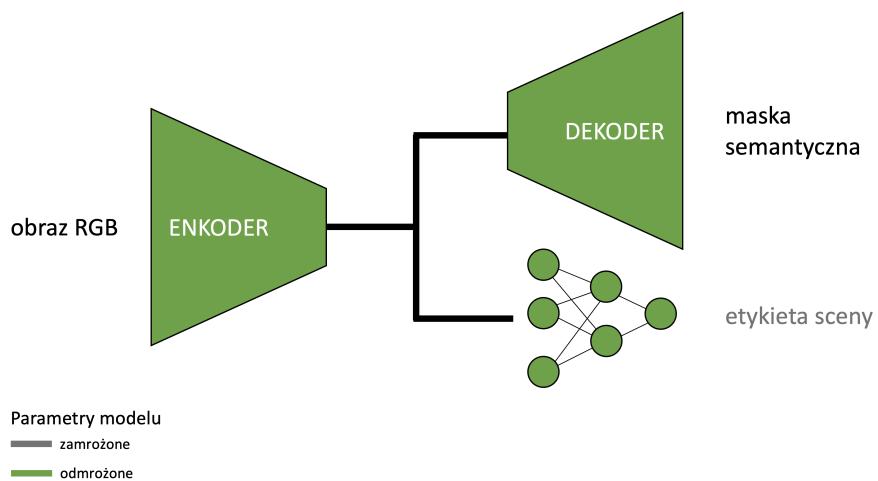
Ostatecznie porównano jakość na przesztrewni obu etapów.



Rysunek 3.3. Architektura sieci zastosowana w pracy inżynierskiej.



Rysunek 3.4. Podejście jednozadaniowe.



4. Eksperymenty

4.1. Zbiór danych

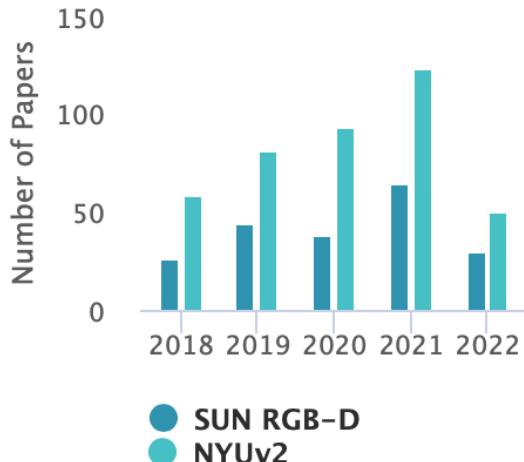
Dane są kluczową częścią głębokiego uczenia. Duży zbiór danych oznaczonych adnotacjami na poziomie pikseli jest potrzebny do wytrenowania wydajnego modelu segmentacji semantycznej. Typowe zestawy danych do segmentacji semantycznej to Cityscapes, PASCAL VOC i ADE20K. Podobnie w przypadku klasyfikacji scen wymagany jest duży zbiór danych z odpowiednią informacją o etykiecie. Popularne zestawy danych do klasyfikacji scen obejmują NYUv2, SUN RGB-D, Matterport3D i ScanNet.

Zbiór danych powinien ściśle odpowiadać założeniom postawionym w pracy. Zatem zbiór danych powinien zawierać kategorie scen, segmentacje obrazów

Po prześledzeniu wielu zbiorów danych udało się sprostać powyższym wymaganiom, uzyskując dwa podobne zbiorów danych - NYUv2 oraz SUN RGBD. Ostatecznie wybrano NYUv2 z uwagi, że zbiór ten został zawiera zdjecia pomieszczeń, w które nie są posprzątane. Fakt ten uznano, za ważny, iż uważało, że będzie przekładał się na lepsze rezultaty w

4. Eksperymenty

naturalnych warunkach. Co więcej NYUv2 jest też częściej cytowany niż SUN RGBD (rys. 4.1).



Rysunek 4.1. Szacowana liczba cytowań w latach 2018-2022 [paperswithcode.com]

4.2. Analiza zbioru danych

Eksploracyjna analiza danych (ang. EDA) to proces eksploracji i zrozumienia cech zbioru danych przed zbudowaniem modelu. Omówione zostanie znaczenie EDA w głębowym uczeniu oraz możliwości wykorzystania do poprawy wydajności i interpretowalności modeli głębokiego uczenia.

Jakość danych

Jednym z głównych powodów, dla których EDA jest ważne w wizji komputerowej, jest to, że może pomóc w identyfikacji problemów ze zbiorem danych, takich jak brakujące wartości, wartości odstające lub nieprawidłowe etykiety, które mogą wpływać na wydajność modelu wizji komputerowej. Przeprowadzając EDA, możemy uzyskać głębsze zrozumienie danych i zidentyfikować wszelkie problemy, które należy rozwiązać przed zbudowaniem modelu.

Wstępne przetwarzanie danych

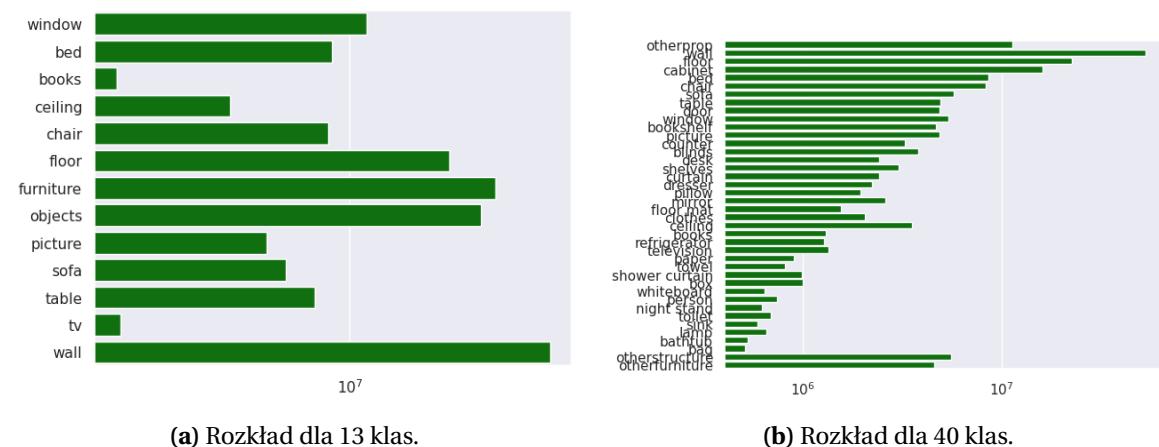
EDA może być również wykorzystana do określenia, które kroki przetwarzania wstępnego (ang. preprocessing), takie jak augmentacja, są niezbędne do poprawy wydajności modelu wizji komputerowej. Badając dane i rozumiejąc ich charakterystykę, jesteśmy w stanie lepiej dostosować różne techniki wstępnego przetwarzania danych.

Identyfikacja tendencyjności

EDA może być również wykorzystana do identyfikacji potencjalnych błędów w zbiorze danych, takich jak skośne rozkłady klas, które mogą wpływać na wydajność modelu widzenia komputerowego i prowadzić do niesprawiedliwych prognoz. Przeprowadzając

EDA, możemy zidentyfikować wszelkie uprzedzenia w danych i podjąć kroki w celu ich rozwiązania przed zbudowaniem modelu.

EDA przeprowadzone na zbiorze NYUv2 dostarczyło wielu interesujących szczegółów. W zbiorze domyślnie znajduje się 795 przykładów trenujących oraz 654 przykładów testujących. Ze zbioru testowego wyodrębniono zbiór walidacyjny stanowiący 20% zbioru testowego. Ponadto sprawdzono rozkład klas na przestrzeni całego zbioru danych. W przypadku zadania segmentacji semantycznej do dyspozycji był wybór 894, 40 lub 13 klas przedmiotów. Im rozróżnialność była większa tym większe okazywały się dysproporcje w rozkładzie. Histogramy dla 13 i 40 klas przedstawiono na rysunku 4.2. Podobna sytuacja miała miejsce dla zadania klasyfikacji z tą różnicą, iż scalanie klas należało dokonać ręcznie. Taki krok był kluczowy, gdyż pierwotny rozkład był silnie zdominowany przez kilka klas. Ostatecznie wybrano 13 klas dla klasyfikacji (rys. 4.3b) oraz scalone 7 dla segmentacji (rys. 4.3b).



Rysunek 4.2. Porównanie rozkładu ilości pixeli dla zadania segmentacji semantycznej.

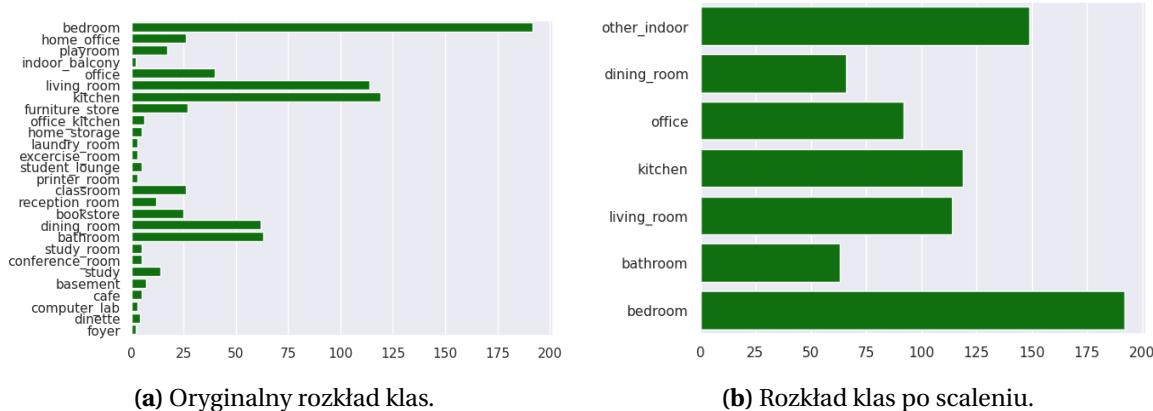
4.3. Opis eksperymentów

Przygotowanie danych

Obrazy RGB zostały poddane normalizacji ze średnią (0.485, 0.456, 0.406) oraz odchyleniem standardowym (0.229, 0.224, 0.225), która odpowiada parametrom rozkładu normalnego na zbiorze ImageNet. Baza ImageNet służyła do wytrenowania enkodera, a więc pierwszej części modelu.

Istnieje kilka różnych technik normalizacji, które mogą być stosowane w problemach z widzeniem komputerowym, takich jak normalizacja min-max i normalizacja rozkładem normalnym. W pracy „Normalization Techniques in Training DNNs: Methodology, Analysis and Application” Lei et. al. [10], autorzy udowadniają, że normalizacja stabilizuje i przyśpiesza trening oraz prawdopodobnie prowadzi do poprawy generalizacji.

4. Eksperymenty



(a) Oryginalny rozkład klas.

(b) Rozkład klas po scaleniu.

Rysunek 4.3. Porównanie rozkładu klas dla zadania klasyfikacji sceny.

Normalizacja jest ważnym krokiem przetwarzania wstępniego w problemach widzenia komputerowego, ponieważ może pomóc w poprawieniu wydajności modelu. Normalizacja odnosi się do procesu skalowania danych wejściowych tak, aby miały w przybliżeniu średnią 0 i odchylenie standardowe 1. Pomaga to zapewnić, że dane wejściowe są w spójnym zakresie i mają podobny rozkład, co może poprawić model. Model

Jako model użyto DeepLabv3, który rozszerzono o dodatkową głowę klasyfikacyjną. Umieszczono ją naturalnie zaraz za enkoderem, a przed dekoderem. Głowa klasyfikacyjna przedstawia się jako sieć w pełni połączona (FC) z dwiema warstwami.

TO TRZEBA ZWIUZALIZOWAĆ!

Listing 1. Struktura głowy klasyfikacyjnej

```
1 nn.AdaptiveAvgPool2d((1, 1)),  
2 nn.Flatten(),  
3 nn.BatchNorm1d(num_filters),  
4 nn.Dropout(p=0.25),  
5 nn.Linear(num_filters, out_features=256, bias=False),  
6 nn.ReLU(inplace=True),  
7 nn.BatchNorm1d(256),  
8 nn.Dropout(p=0.25),  
9 nn.Linear(in_features=256, out_features=scene_classes, bias=False),
```

Funkcja straty

W obu przypadkach jako funkcję straty wykorzystano ważoną entropię skrośną. Wagi odzwierciedlały odwrotność liczności w zbiorze. Dla klasyfikacji liczona była ilość klas, natomiast dla segmentacji ilość pikseli.

Uczenie

ssssssssss

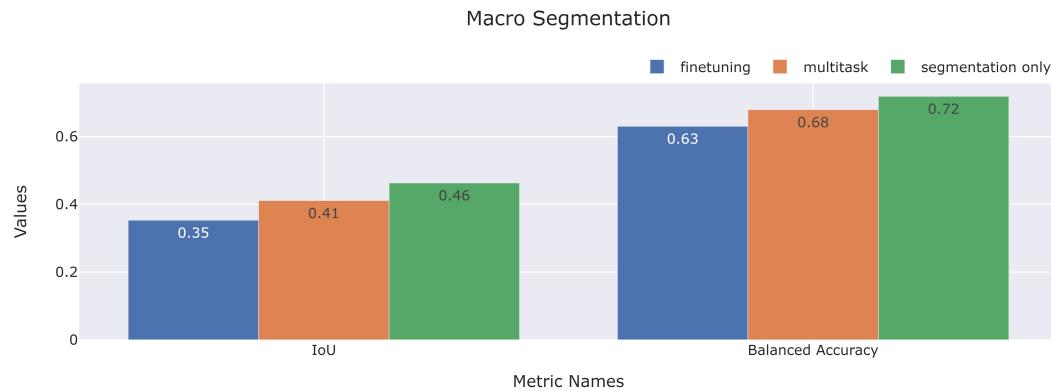
4.4. Wyniki

W tym rozdziale zostaną przedstawione empiryczne wyniki badań nad wspólną segmentacją semantyczną i klasyfikacją sceny w środowiskach wewnętrznych. Badania mają na celu opracowanie i ocenę różnych znanych i aktualnych technik uczenia głębokich sieci neuronowych. Aby to osiągnąć, przeprowadzono serię eksperymentów na zbiorze na reprezentacyjnych zbiorach danych. Analiza dotyczyła zarówno miar jakości sensu stricto jak i miar wydajnościowych proponowanych metod. Rozważano różne metryki oceny, takich jak ogólna dokładność, indeks Jaccarda znany w literaturze jako intersection over union (IoU), miara F1 i wydajność obliczeniowa. Wyniki uzyskane w tym rozdziale zapewniają cenny wgląd w mocne strony i ograniczenia proponowanych metod.

W pierwszej kolejności metody zostaną zbadane pod względem wymienionych wcześniej miar jakości w postaci ogólnej - niezagregowanej, osobno dla segmentacji oraz klasyfikacji. Omawiane metryki należy rozumieć jako średnia miara jakości na każdej z klas, a więc makrośrednie. Makrośrednie metryki są stosowane przy ocenie wydajności algorytmów dla zadań takich jak segmentacja semantyczna i klasyfikacja sceny, ponieważ zapewniają bardziej wszechstronną ocenę ogólnej jakości algorytmu. Metryki makrośrednie uwzględniają wydajność algorytmu na wszystkich klasach obiektów i regionów w obrębie sceny, a nie tylko koncentrują się na jakości na najbardziej powszechnych lub najłatwiejszych do sklasyfikowania klasach. W przypadku stosowania metryki makrośredniej jakość dla każdej klasy jest obliczana oddzielnie, a ogólna jakość jest obliczana jako średnia jakości poszczególnych klas. Stanowi to kontrast do metryki mikrośredniej, która oblicza ogólną jakość poprzez zsumowanie całkowitej liczby wyników dla wszystkich klas. Użycie makrośrednich metryk może być szczególnie ważne w scenariuszach, w których liczba instancji każdej klasy jest niezrównoważona lub gdy istnieje duża liczba klas. W takich przypadkach, mikrośrednie metryki mogą być mylące, ponieważ mogą być pod silnym wpływem najbardziej powszechnych klas, podczas gdy zaniedbują te mniej powszechnne. Zatem makro analiza pokaże generalne rezultaty oraz otworzy dyskusję do dalszych, bardziej połączonych badań na rozważanym problemem.

Rozpoczynając od segmentacji rozważamy 3 scenariusze testowe. Pierwszym z nich jest uczenie wyłącznie klasyfikacji rozumianej jako uczenie enkodera i sieci segmentacyjnej z pominięciem części klasyfikacyjnej. Pozwoli to odpowiedzieć na pytanie czy bardziej zaawansowane techniki uczenia polepszają, a może pogorszą działanie modelu. Drugim scenariuszem jest uczenie wielozadaniowe, gdzie cały model jest odmrożony, a błąd jest propagowany zarówno przez segmentację jak i klasyfikację. Ostatnim eksperymentem jest sprawdzenie technik transferu wiedzy, a szczególnie tak zwanego finetunowania. Model w pierwszym etapie uczy się przy zamrożonym enkoderze, dopiero na koniec jest odmrażany w celu dostrojenia wyników.

Analizując rysunek 4.4 nie trudno zauważać, że najlepsze rezultaty otrzymano w dla



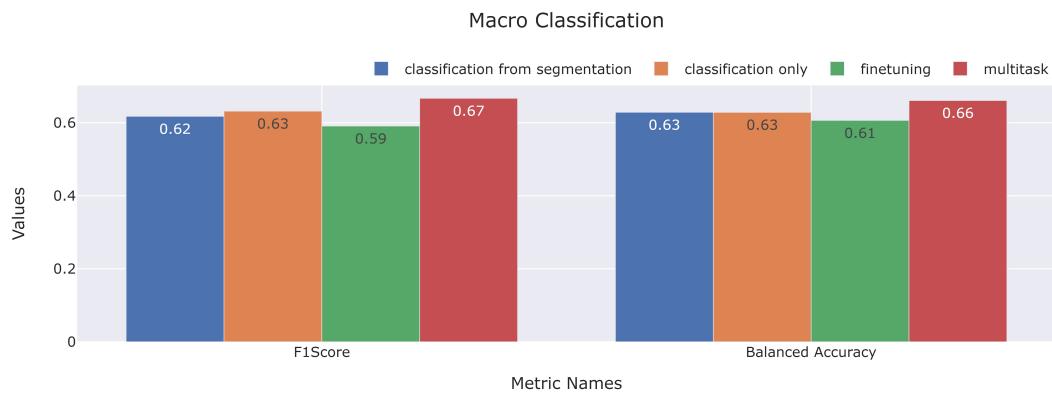
Rysunek 4.4. Porównanie miar IoU oraz dokładności dla segmentacji sceny.

uczenia wyłącznie segmentacji. Kolejnym wynikiem jest uczenie wielozadaniowe. Jako najsłabsze podejście okazuje się metoda finetunowania. Widać, że relacje jakości są zachowane dla każdej z metryk, a więc zarówno dla mocy IoU jak i zbalansowanej dokładności (bAcc). Widać, że miara IoU wypada gorzej niż bAcc. Wyniką mogą sugerować, że trudno jest przeprowadzić transfer wiedzy z ImageNetu, gdyż finetunowanie wypada najsłabiej. Jest to naprawdopodobniej spowodowane zupełnie innym rozkładem klas dla wspomnianej bazie. Analiza sceny w przeciwnieństwie do klasyfikacji najczęściej cechuje się długogonowym rozkładem klas. Drugim istonym szczegółem jest fakt, iż wagi dekoderu i głowy segmentacyjnej są losowe. Uczenie wielozadaniowe zgodnie z zakładanymi wynikami nie polepsza segmentacji, gdyż łączna przestrzeń segmentacji i klasyfikacji jest niewątpliwie bardziej trudniejsza do optymalizacji.

Przechodząc do klasyfikacji wyróżniamy 4 scenariusze testowe. Pierwszym jest uczenie wyłącznie klasyfikacji, analogicznie jak wyżej, a więc przy wyłączonej części segmentacyjnej. Kolejnymi są wspomniane wcześniej uczenie wielozadaniowe oraz finetuning. Nowym scenariuszem jest skorzystanie z wytrenowanej wcześniej wyłącznie segmentacji, a następnie wyłączenie wszystkiego poza siecią gęstą.

Rezultaty przedstawia rysunek 4.5. Od razu da się zauważyć, że wyniki cechują mniejsze odchylenie standardowe oraz, analizując łącznie mierę F1 oraz zbalansowaną dokładność, średnia. Fakt ten jest prawdopodobnie wynikiem znacznie mniejszej ilości parametrów uczących. Jako najlepszy rezultat uzyskuje uczenie wielozadaniowe. Ciekawym wydaje się fakt, że uczenie wyłącznie klasyfikacji jest słabsze w tym przypadku. Prawdopodobnie poprzez uczenie wielozadaniowe enkoder wygenerował lepszą przestrzeń reprezentacji, co bezpośrednio wpływa na klasyfikację sceny.

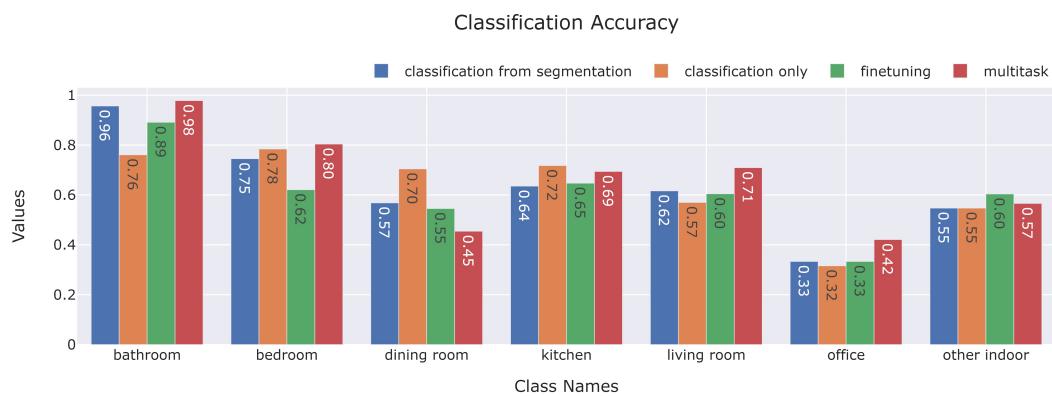
Analizowanie jakości algorytmu dla każdej z klas osobno jest ważne, ponieważ pozwala na bardziej szczegółowe zrozumienie mocnych i słabych stron algorytmu. Rozważając



Rysunek 4.5. Porównanie miar F1 oraz dokładności dla klasyfikacji sceny.

ogólną jakość algorytmu przy użyciu metryki makrośredniej, nie jest od razu jasne, w których klasach algorytm radzi sobie dobrze, a z którymi ma problemy. Analizując jakość każdej klasy osobno, można zidentyfikować konkretne klasy, z którymi algorytm ma problemy i podjąć kroki w celu poprawy wydajności w tych klasach.

Rysunek 4.6 przedstawia dokładność dla każdej z klas dla zadania klasyfikacji sceny. Trudno jednoznacznie określić która metoda sprawdza się tutaj najlepiej. Uczenie wielozadaniowe wypada najlepiej dla klas: łazienka, pokój dzienny, salon, biuro. Uczenie wyłączeni klasyfikacji jest najlepsze dla klas jadalnia oraz kuchnia. W pozostałych przypadkach klasa inne pomieszczenia jest najlepiej wykrywana przez scenariusz finetunowania. Uczenie klasyfikacji z segmentacji nigdy nie osiąga najlepszego wyniku. Biorąc pod uwagę miarę F1

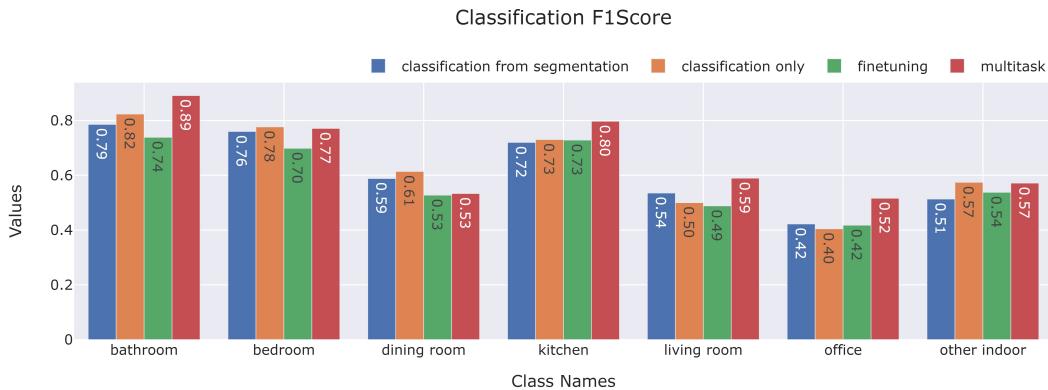


Rysunek 4.6. Porównanie dokładności klasyfikacji sceny z rozróżnieniem konkretnych klas.

(rys.4.7) również nie jesteśmy w stanie wyróżnić faworyzowanej metody. W porównaniu z wcześniej analizowaną dokładnością widać, że uczenie wielozadaniowe utrzymuje w

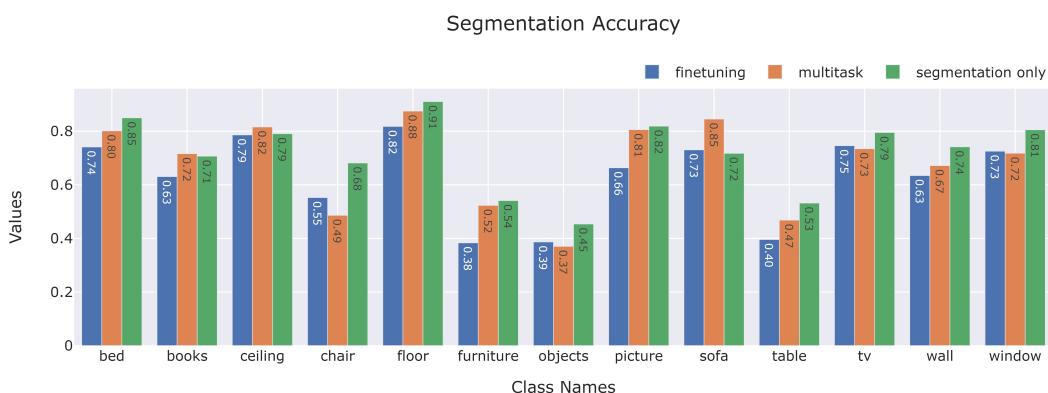
4. Eksperymenty

większości przypadku bardzo dobre rezultaty. Widać też, że wyniki w obrębie każdej z klas mało różnią się między sobą.



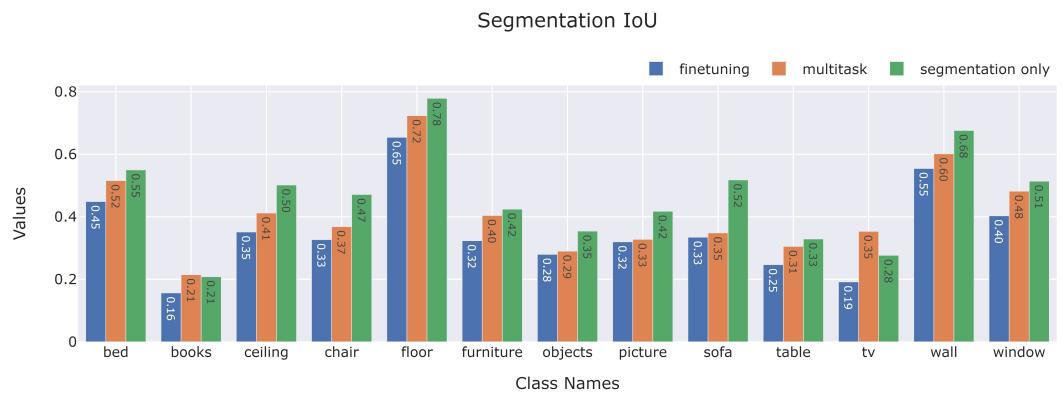
Rysunek 4.7. Porównanie miary F1 dla klasyfikacji sceny z rozróżnieniem konkretnych klas.

Analizując rysunek 4.8 przestawiający dokładność w zadaniu segmentacji semantycznej, widać, że niektóre z zadań wypadają znacznie gorzej niż pozostałe. Sytuacja ta dotyczy klas meble, stół, obiekty. Uczenie wyłącznie segmentacji okazało się najlepsze dla klas łóżko, podłoga, meble, obiekty, obraz, tv, ściana oraz okno. Stanowi to ponad połowę wszystkich możliwych klas. Uczenie wielozadaniowe uzyskało najlepsze wyniki dla klas książki, sufit, sofa. Przypadek funetunowania nigdy nie osiągnął najlepszego rezultatu.



Rysunek 4.8. Porównanie dokładności segmentacji z rozróżnieniem konkretnych klas.

Na rysunku 4.9 przedstawiono IoU dla segmentacji semantycznej. Widać tutaj dużą dysproporcję między klasami podłoga, ściana, a pozostałymi klasami. Jest to zrozumiałe, klas te występują stosunkowo często na obrazie. Uczenie wyłącznie segmentacji uzyskuje najlepsze wyniki na wszystkich klasach z wyłączeniem książek oraz telewizorów. W tych przypadkach najlepsze okazuje się uczenie wielozadaniowe



Rysunek 4.9. Porównanie miary IoU segmentacji z rozróżnieniem konkretnych klas.

5. Podsumowanie

Bibliografia

- [1] J. Long, E. Shelhamer i T. Darrell, “Fully convolutional networks for semantic segmentation”, w *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, s. 3431–3440.
- [2] O. Ronneberger, P. Fischer i T. Brox, “U-net: Convolutional networks for biomedical image segmentation”, w *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, s. 234–241.
- [3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy i A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, t. 40, nr. 4, s. 834–848, 2018. DOI: 10.1109/TPAMI.2017.2699184.
- [4] D. Zeng, M. Liao, M. Tavakolian, Y. Guo, B. Zhou, D. Hu, M. Pietikäinen i L. Liu, “Deep learning for scene classification: A survey”, *arXiv preprint arXiv:2101.10531*, 2021.
- [5] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi i A. Agrawal, “Context encoding for semantic segmentation”, w *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, s. 7151–7160.
- [6] S. Ruder, “An overview of multi-task learning in deep neural networks”, *arXiv preprint arXiv:1706.05098*, 2017.
- [7] R. Caruana, “Multitask learning: A knowledge-based source of inductive bias”, w *Proceedings of the Tenth International Conference on Machine Learning*, Citeseer, 1993, s. 41–48.
- [8] S. Iizuka, E. Simo-Serra i H. Ishikawa, “Let there be color! Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification”, *ACM Transactions on Graphics (ToG)*, t. 35, nr. 4, s. 1–11, 2016.
- [9] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore i L. Shapiro, “Y-Net: joint segmentation and classification for diagnosis of breast biopsy images”, w *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, s. 893–901.
- [10] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu i L. Shao, “Normalization techniques in training dnns: Methodology, analysis and application”, *arXiv preprint arXiv:2009.12836*, 2020.

Spis rysunków

2.1 Problem różnorodności wewnątrzklasowej oraz wieloznaczności semantycznej [4].	15
2.2 Segmentacja wewnątrz pomieszczeń [5].	16
3.1 Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification 2016 [8].	17
3.2 Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images 2018 [9].	17
3.3 Architektura sieci zastosowana w pracy inżynierskiej.	18
3.5 Architektura sieci jako uczenie wielozadaniowego.	19
4.4 Porównanie miar Iou oraz dokładności dla segmentacji sceny.	24
4.5 Porównanie miar F1 oraz dokładności dla klasyfikacji sceny.	25
4.6 Porównanie dokładności klasyfikacji sceny z rozróżnieniem konkretnych klas. . . .	25
4.7 Porównanie miary F1 dla klasyfikacji sceny z rozróżnieniem konkretnych klas. . . .	26
4.8 Porównanie dokładności segmentacji z rozróżnieniem konkretnych klas.	26
4.9 Porównanie miary IoU segmentacji z rozróżnieniem konkretnych klas.	27

Spis tabel