

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. finetuning	16
2.6. Uczenie wielozadaniowe	16
3. Rozwiązanie	18
3.1. Przegląd rozwiązań	18
3.2. Rozwiązywanie problemu	22
3.2.1. Uczenie wielozadaniowe	22
3.2.2. Wyłącznie klasyfikacja	22
3.2.3. Wyłącznie segmentacja	23
3.2.4. Finetuning	23
3.2.5. Równoległa klasyfikacja z segmentacją	23
3.2.6. Szeregową klasyfikacją z segmentacją	24
4. Eksperymenty	25
4.1. Zbiór danych	25
4.2. Analiza zbioru danych	25
4.3. Opis eksperymentów	26
4.4. Wyniki	28
4.4.1. Analiza miar jakości	28
4.4.2. Analiza czasowa	32
4.5. Porównanie jakości	33
4.5.1. Segmentacja semantyczna	33
4.5.2. Klasyfikacja sceny	41
5. Podsumowanie	44
Bibliografia	45
Spis rysunków	48
Spis tabel	48

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].

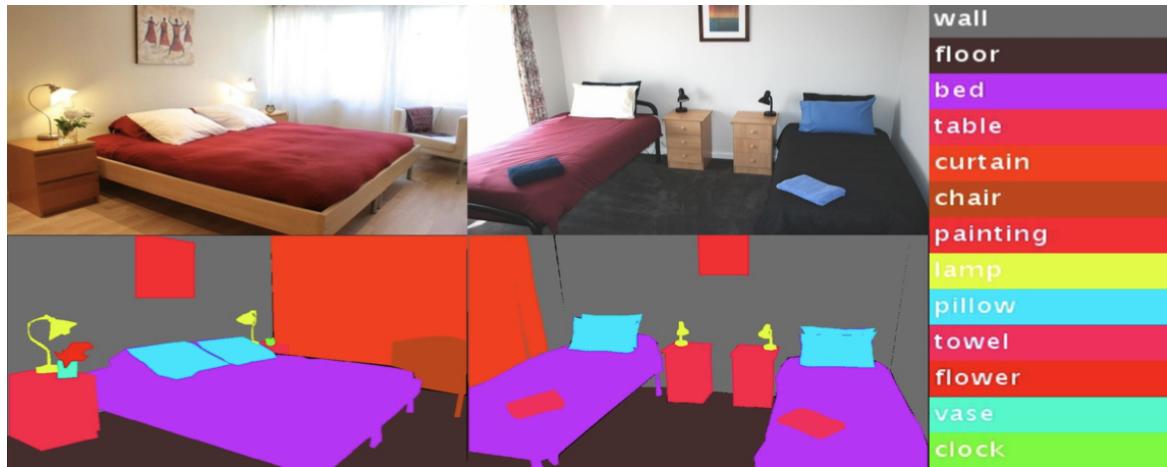
planowego, np. czy na obrazie znajduje się pies, czy kot. Klasyfikacja sceny natomiast musi wziąć pod uwagę wszystkie cechy obrazu, zarówno tła, jak i pierwszego planu, by określić odpowiednie miejsce.

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. finetuning

XXXXXXXXXXXXXX

2.6. 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 zwięźle 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 dwa główne podejścia do uczenia wielozadaniowego - twarde oraz miękkie dzielenie wag sieci (soft/hard parameter sharing). Twarde dzielenie wag jest najczęściej stosowane. Polega na uwspólnieniu pierwszej części sieci, odpowiedzialnej za zdefiniowanie przestrzeni reprezentacji (ang. backbone) oraz rozdzieleniu kolejnych warstw związanych z konkretnym zadaniem. Miękkie dzielenie wag polega na zbudowaniu wielu sieci, odpowiednich dla danego zadania. Co więcej, sieci te podczas uczenia są regularyzowane w ten sposób, aby zachęcić je do posiadania jak najbardziej różnych wag.

Takie podejścia mają prawo działać jedynie w przypadku, kiedy dwa zadania są powiązane ze sobą. Powstało wiele prac poświęconych odpowiedzi na pytanie, które zadania warto wybrać, a które należy rozpatrywać osobno. Jednym z takich dzieł jest praca zespołu ze Stanfordu „Which Tasks Should Be Learned Together in Multi-task Learning?” Standley et. al. (2020) [8]. Przedstawia ona pojęcie negatywnego wpływu (ang. negative transfer), który najprościej rzecz ujmując sprawia, że sieć uczy się gorzej niż pojedyncze sieci. Autorzy zbadali, że największy wpływ na jakość uczenia wielozadaniowego ma właśnie odpowiedni dobór zadań, a niekoniecznie rozmiar zbioru danych czy wielkość modelu. Oczywiście należy zwrócić uwagę, że przytoczone czynniki nie są bez znaczenia, jedynie w przypadku doboru zadań mają pomijalne znaczenie. Co ciekawe zadania aficzne względem siebie mogą mieć dodatni wpływ w przypadku transferu wiedzy, a nie muszą być aficzne w kontekście uczenia wielozadaniowego.

Gdy jednak zadania są pokrewne względem siebie jesteśmy w stanie zaobserwować konkretne korzyści związane ze wspólnym uczeniem. Ruder wymienia kilka najważniejszych. Po pierwsze zyskujemy tak zwaną bezpośrednią augmentację danych (ang. implicit data augmentation). Każde z zadań posiada pewien szum związany z konkretnym zadaniem. Uczenie wielu zadań pozwala w pewnym stopniu wyeliminować szum związany z konkretnym zadaniem na rzecz lepsze generalizacji. Kolejną zaletą jest lepsze skupienie uwagi na ważnych informacjach. Ma to szczególny znaczenie w przypadku gdy dane są oganiczone lub wielowymiarowe. Uczenie wielozadaniowe może pomóc w wyborze tych najbardziej znaczących cech. Co więcej, wspólna wiedza zdobyta podczas uczenia może okazać się znacząca. Niktore cechy są łatwiejsze do wykrycia dla jednego zadania, inne dla drugiego. Łacząc te informacje przez tak zwane „podsłuchiwanie” (ang. eavesdropping) model jest w stanie zbudować lepszą przestrzeń reprezentacji. Oprócz zyskania na jakości modelu, przypadek twardego dzielenia wag pozwala znaczco ograniczyć wielkość modelu. Nie trzeba bowiem stosować wielu backbone’ów, które stanowią największą część modelu w kontekście liczby parametrów. Implikuje to znacznie zmniejszenie czasu uczenia oraz wnioskowania [8].

NAPISAĆ PODSUMOWANIE

3. Rozwiązanie

Zadania wizji komputerowej mogą zostać podjęta na wiele sposobów. Szczególnie interesujące są podejścia do zadania łącznej segmentacji semantycznej i klasyfikacji sceny we wnętrzach. W tym rozdziale przedstawione zostaną wybrane metody, które zostały sprawdzone w ramach analizy problemu. W swoich rozważaniach będę bezpośrednio odnosił się do pytań badawczych postawionych w celu pracy, a więc:

- 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?

Opis rozwiązań problemu zostanie poprzedzony przeglądem rozwiązań. Analiza dotychczasowych pozwoli lepiej ukierunkować badania. Korzystając z doświadczenia innych, będzie można wyrobić sobie intuicję, która pomoże podejmować konkretne decyzje.

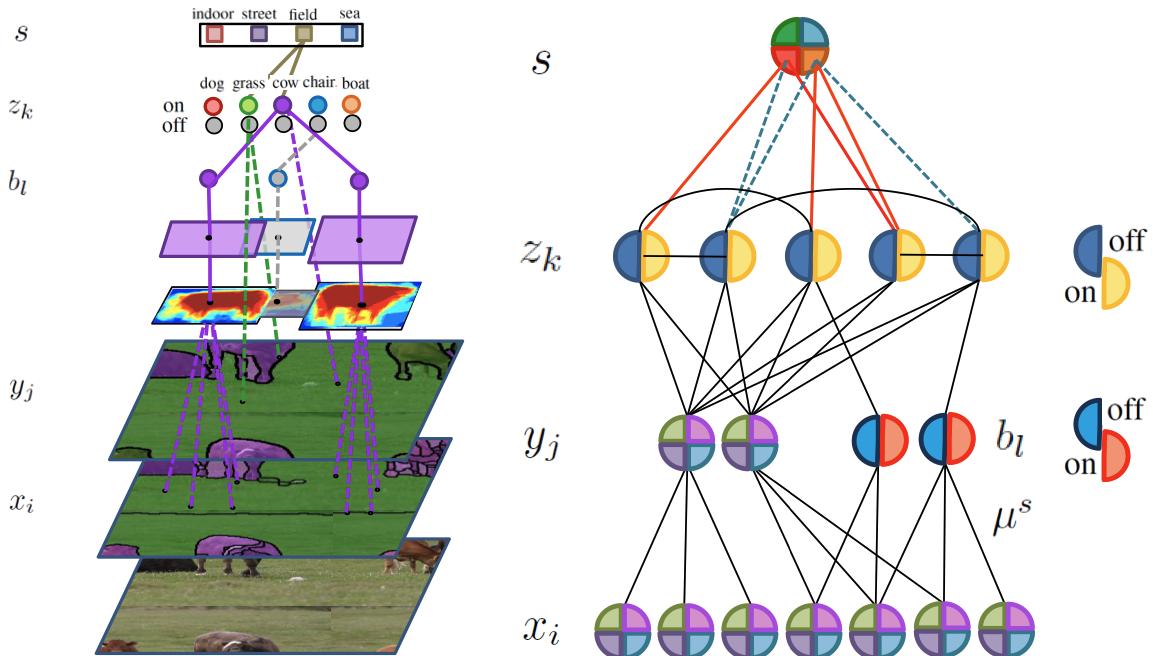
3.1. Przegląd rozwiązań

Przegląd literatury jest kluczowym aspektem każdej pracy naukowej. W tym rozdziale zostaną przedstawione wyłącznie rozwiązania obejmujące łączną segmentację semantyczną oraz klasyfikację sceny. Szczególny nacisk położony zostanie na architektury głębokich sieci neuronowych z dogłębną analizą przepływu inferencji przez nie. Niestety

przyjęte założenia w pracy nie zostały opisane przez nikogo wcześniej, zgodnie z najlepszą wiedzą autora. Niektóre prace naukowe przedstawiają ten sam problem to jest klasyfikacji i segmentacji łącznie, ale obejmują go w innej domenie danych. Z drugiej artykuły obejmujące środowiska wnętrza są dobrze zdefiniowane, jednak często w swoich rozwiązań korzystają z obrazu głębki, który nie zawiera się w zakresie badań tej pracy. Nie mniej wszystkie poniższe artykuły stanowią cenne źródło informacji, które należy mniej lub bardziej dostosować do rozważanego problemu.

Pierwszym z prezentowanych artykułów jest „Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation” autorstwa Yao j. et al (2012)[9]. Prezentuje on algorytm, który ówcześnie wyznaczył najlepsze podejście (ang. state-of-the-art (SOTA)). Autorzy wskazują tutaj, że połączenie rozważanych zadań okazało się owocne nie tylko pod względem jakości, ale również wydajności w kontekście czasowym. Yao J. et al zwracają uwagę na połączenie szeregowe, które niestety propaguje

błąd w kolejnych zadaniach, a było dotychczasowo szeroko stosowane. W swojej pracy wykorzystują podejście równolegle zgodne z rysunkiem 3.1. Podsumowując „Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation” nie jest propozycją architektury głębskiej sieci. Wskazuje on na problemy z łączeniem, zadań szeregowo, jednocześnie udowadniając, że taka praktyka był ówcześnie stosowana, więc nie można uznawać stosowania połączenia szeregowego jako niedopuszczalnego.

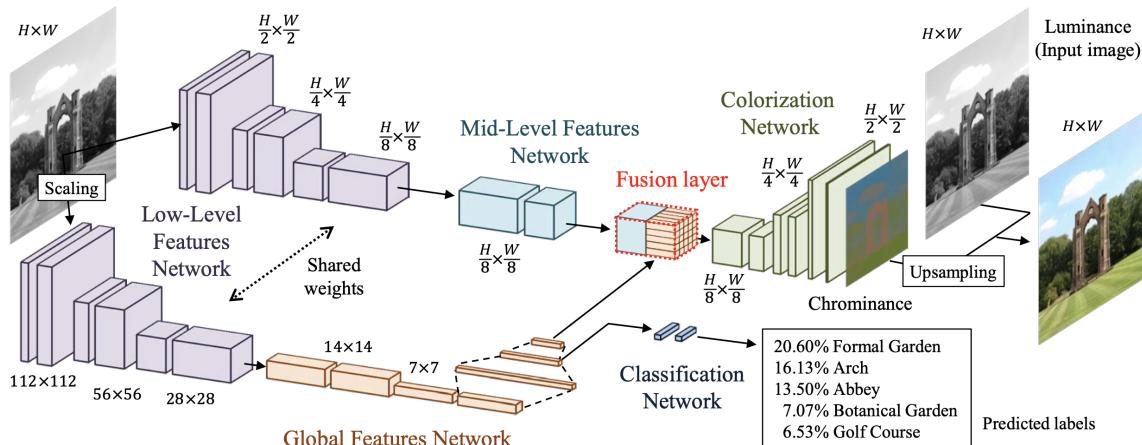


Rysunek 3.1. Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation (2012) [9].

Artykuł „Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification 2016 [10].” (2016) przedstawia rozwiązanie problemu jednoczesnego klasyfikowania sceny oraz kolorowania zdjęć. Do realizacji zadania kolorowania potrzebna jest sematyczna maska. Wynika z tego, że kolorowanie jest rozszerzeniem segmentacji semantycznej. Rozumiejąc towarzyszące analogie można przejść do analizy rozwiązania. Przedstawiona architektura (rys.3.2) jest przykładem sieci wielozadaniowej, używającej miękkiego dzielenia parametrów, ale tylko i wyłącznie w obrębie pierwszej części sieci. Szczególnie ciekawa jest konkatencja cech wysokiego poziomu (Global Features Network) z cechami średnio-poziomowymi (Mid-Level Features Network), która ma miejsce w warstwie fuzji (Fusion layer). Iizuka et al. formułują wniosek oznajmiający o kluczowym znaczeniu tej warstwy w kontekście całego zadania. Wiedza o scenie zdjęcia może dostarczyć informacji wpływających na decyzję, czy na obrazie znajduje się niebo czy trawa. Rozważając sceny wewnętrz oczywiste jest, że nie będzie tam takich grup semantycznych. Podsumowując,

3. Rozwiązanie

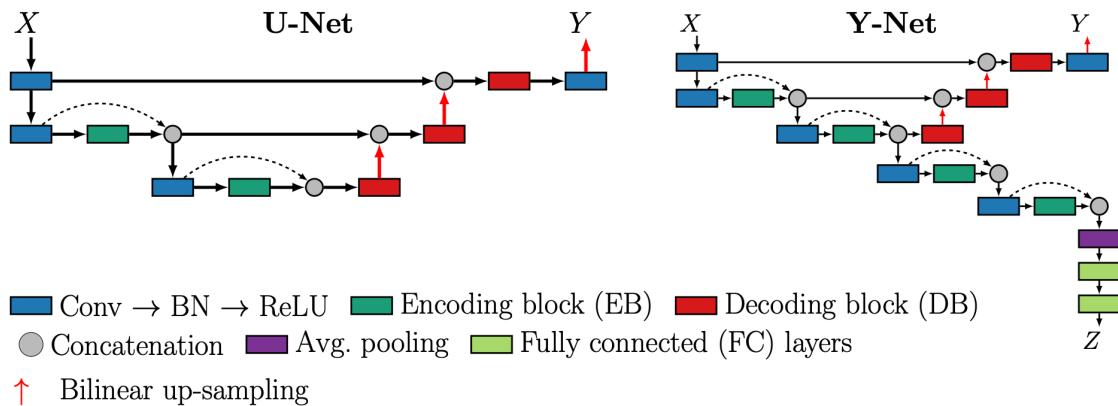
cechy nauczone na zadaniach klasyfikacji i segmentacji, mogą wzajemnie pozytywnie na siebie wpływać, realizując pozytywny transfer.



Rysunek 3.2. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification 2016 [10].

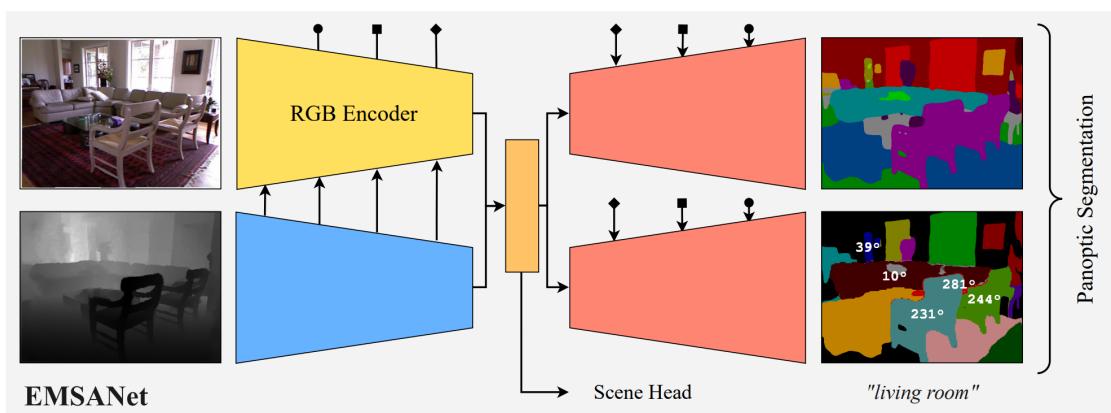
Zastosowanie łącznej segmentacji oraz klasyfikacji tym razem w domenie medycznej przedstawia „Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images” [11] (2018). Zadanie te są realizowane przez twardé dzielenie parametrów w kontekście uczenia wielozadaniowego (rys.3.3). Architektura jest prostym rozszerzeniem klasycznego U-Netu. Autorzy wskazują, że taki zabieg powodują dużą modulatność, ponieważ do dowolnego modelu segmentacji można podłączyć sieć klasyfikacyjną. Przeprowadzone eksperymenty dla segmentacji udowodniły, że dokładność pozostała na tym samym poziomie. W przypadku klasyfikacji wyniki były wyższe niż dotychczasowe SOTA na tym zbiorze. Jako funkcję straty autorzy użyli sumę entropii skrośnej każdego z zadań. Podsumowując zadanie zadanie segmentacji osiągnęło ten sam wysoki wynik co SOTA, a zadanie klasyfikacji ustanowiło nowe SOTA na tym zbiorze ucząć się znacznie mniej parametrów.

Najbliższy artykuł tej pracy inżynierskiej jest „Efficient Multi-Task RGB-D Scene Analysis for Indoor Environments” [12] (2022), który został opublikowany w czasie tworzenia tej pracy. Przedstawia on jedną głęboką sieć neuronową rozwiązującą następujące zadania: segmentacja semnatyczna oraz segmentacja instacji (łącznie ang. panoptic segmentation), estymacja orientacji instacji oraz klasyfikację sceny. Rozważaną przez autorów domeną są podobnie jak w przypadku tej pracy sceny wnętrz. Znaczą różnicą poza dodatkowymi zadaniami jest użycie przez zespołu Seichter et al. informacji o głębi. Zgodnie z wnioskami z nieniejszego artykułu przetwarzanie łączne obrazów RGB i głębi jest kluczowe z punktu widzenia jakości predykcji. Każde z zadań zostało na początku trenowane osobno by ustalić punkt odniesienia. Architektura jest przedstawiona na rysunku 3.4. Autorzy zdecydowali się na twardé dzielenie parametrów, argumentując całkowitą niezależnością w przypadku chęci wyłączenia jednego z zadań z wnioskowania. Trening każdej sieci z



Rysunek 3.3. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images 2018 [11].

osobna był rozważany pod względem wielu backbone'ów ze zróżnicowaniem na uczenie wyłącznie obrazu glebi, obrazu RGB lub RGB-D. Generalnie w przypadku segmentacji oraz klasyfikacji większy backbone wpływał na polepszenie wyników. Trenując zadania łącznie zdecydowano się na ważoną sumę entropii skrośnej dla zadania segmentacji i klasyfikacji w proporcjach odpowiednio 3:1. Przyjęty krok uczenia, będąc sprawdzonym przez przeszukiwanie liniowe (ang. grid search), jest wyjątkowo duży, bo wynosi 0.02. Autorzy zastosowali zaawansowane techniki dostosowywania kroku czenia w trakcie treningu poprzez użycie planista polityki jednego cyklu (ang. one cycle policy scheduler). Jako optymalizator użyto SGD z momentem oraz drobną regularyzacją. Podsumowując, zgodnie z prezentowanymi wynikami na wspólnej segmentacji oraz klasyfikacji autorom nie udało się polepszyć działania modelu na segmentacji semantycznej. Z powodzeniem jednak wzrosła dokładność klasyfikacji na zbiorze NYUv2.



Rysunek 3.4. Efficient Multi-Task RGB-D Scene Analysis for Indoor Environments [12]

3.2. Rozwiązanie problemu

W tym rozdziale zostaną przedstawione eksperymenty, które wykonano w celu zbadania uczenia wielozadaniowego segmentacji semantycznej oraz klasyfikacji sceny w domenie pomieszczeń. Pierwszym założeniem jakiego dokonano było wyznaczenie punktu odniesienia. Z punktu widzenia pracy nałatwiej byłoby znaleźć gotowe wyniki segmentacji oraz klasyfikacji sceny na wybranym zbiorze danych. Niestety żadne z przytaczanych rozwiązań nie odpowiada w pełni zakresowi pracy. Postanowiono stworzyć taki punkt odniesienia samemu przez analogiczne trenowanie sieci segmentacyjnej oraz klasyfikacyjnej osobno.

Posiadając taką wiedzę eksperymentowano dalej z różnymi architekturami uczenia wielozadaniowego. Wybrano uczenie łączne o twardym dzieleniu parametrów. Podejście to ma wiele zalet. [11] podkreśla łatwość i wszechstronność implementacji. Wystarczy dołączyć do modelu część klasyfikacyjną. Co więcej wszyscy autorzy ([11], [12]) chwalą znacznie mniejszą ilość parametrów sieci co bezpośrednio wpływa na czas treningu oraz wnioskowania. Architektura sieci przedstawia się następująco 3.5. Jest to DeepLabv3 rozszerzony za enkoderem o sieć klasyfikacyjną podobnie jak w artykule [11], gdzie rozszerzono sieć U-Net.



Rysunek 3.5. Architektura wielozadaniowej sieci.

3.2.1. Uczenie wielozadaniowe

Uczenie wielozadaniowe zostało zrealizowane przez architekturę z rysunku 3.5. Trening polegał na aktualizowaniu wag całego dostępnego modelu zgodnie zgodnie z propagą wsteczną zagregowanej funkcji straty λ . Zaimplementowano ją jako sumę funkcji strat na każdym z zadań, tak jak w przypadku [11]. Nie stosowano ważenia zadań [12].

$$\lambda = \lambda_{segmentacja} + \lambda_{klasyfikacja}$$

3.2.2. Wyłącznie klasyfikacja

W celu określenia punktu odniesienia wytrenowano model zapominając o podsieci do wyznaczania segmentacji semantycznej. Technicznie skorzystano z modelu wielozadaniowego.

wego. Parametry modułów architektury takie jak dekoder oraz główna segmentacyjna zostały zamrożone oraz nie zostały podawane optymalizatorowi w trakcie treningu. Funkcja straty λ została ograniczona wyłącznie do straty na klasyfikacji poprzez wyzerowanie w każdym kroku straty na segmentacji

$$\begin{aligned}\lambda &= \lambda_{segmentacja} + \lambda_{klasyfikacja} \\ \lambda_{segmentacja} &= 0\end{aligned}$$

3.2.3. Wyłącznie segmentacja

Analogicznie jak w przypadku klasyfikacji należało określić punkt odniesienia również w przypadku segmentacji. Procedura była taka sama jak w przypadku klasyfikacji. Model wielozadaniowy zamrożono w części klasyfikacyjnej oraz wyłączono zamrożone parametry z optymalizacji. Funkcja straty λ została przedstawiona jako

$$\begin{aligned}\lambda &= \lambda_{segmentacja} + \lambda_{klasyfikacja} \\ \lambda_{klasyfikacja} &= 0\end{aligned}$$

3.2.4. Finetuning

Znaną techniką transferu wiedzy jest finetuning. W tym przypadku skorzystano z wytrenowanego enkodera ResNet wytrenowanego na dużej bazie ImageNet. Uczenie przebiegało w dwóch fazach. W pierwszej zamrożono enkoder i starano się osiągnąć jak najlepsze rezultaty dysponując podsiecią klasyfikacyjną i segmentacyjną. Wynika z tego, że pierwszy etap to ni innego jak uczenie wielozadaniowe ale z wyłączonym enkoderem. Dopiero w drugim etapie odmrażany jest również enkoder. Sytuacja wtedy przypomina wcześniejszej omawiane uczenie wielozadaniowe. Jendakże, kluczowy jest dobór hiperparametrów. W pierwszym etapie uczenie przebiega z pewnym krokiem uczenia. W drugim zaś krok uczenia jest znacznie mniejszy.

$$\lambda = \lambda_{segmentacja} + \lambda_{klasyfikacja}$$

3.2.5. Równoległa klasyfikacja z segmentacją

Podejście transferu wiedzy można lekko zmodyfikować. Skorzystano z wcześniejszych przygotowanych wag będących wynikiem wcześniejszej wspomnianej wyłącznie segmentacji. Zamrożono enkoder oraz podsieć segmentacyjną oraz wyłączono te parametry z optymalizacji. Następnie dysponując samą podsiecią klasyfikacyjną przeprowadzono trening.

3. Rozwiązanie

Funckja straty była następująca:

$$\lambda = \lambda_{segmentacja} + \lambda_{klasyfikacja}$$
$$\lambda_{segmentacja} = 0$$

3.2.6. Szeregową klasyfikacją z segmentacją

. Rozwiązaniem odbiegającym od reszty jest przeprowadzenie szeregowej klasyfikacji z segmentacją. Architektura przedstawia się zgodnie z rysunkiem 3.6. W tym eksperymencie sprawdzono jak można skorzystać zgotowych predykcji dotyczących segmentacji. Model aż do głowy segmentacyjnej włącznie został zamknięty oraz wyłączony z optymalizacji. Zmieniają się tylko wagi części klasyfikacyjnej.

$$\lambda = \lambda_{segmentacja} + \lambda_{klasyfikacja}$$



Rysunek 3.6. Arhitekura sieci szeregowej.

W celu łatwej oraz dokładniej ewaluacji

W celu lepsze ewaluacji uczenia wielozadaniowego zdecydowano uściślij wszystkie parametry sieci takie jak architektura jest ta sama zeby dało się porównać nie

nie ma sensu wszystkich scenariusz bo tak

opisać że nie ma sensu porównywać

Poza uczeniem łącznym zbadano też inne znane techniki uczenia jak finetuning.

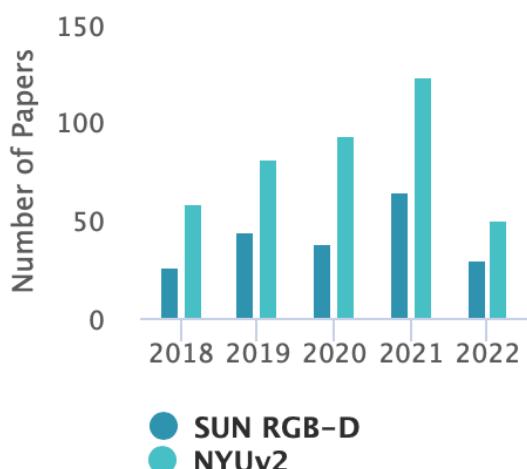
4. Eksperymenty

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 sceny 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 zdjęcia pomieszczeń, w które nie są posprzątane. Fakt ten uznano, za ważny, iż uważało, że będzie przekładał się na lepsze rezultaty w 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 głębokim uczeniu oraz możliwości wykorzystania do poprawy wydajności i interpretowalności modeli głębokiego uczenia.

Jakość danych

4. Eksperymenty

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.

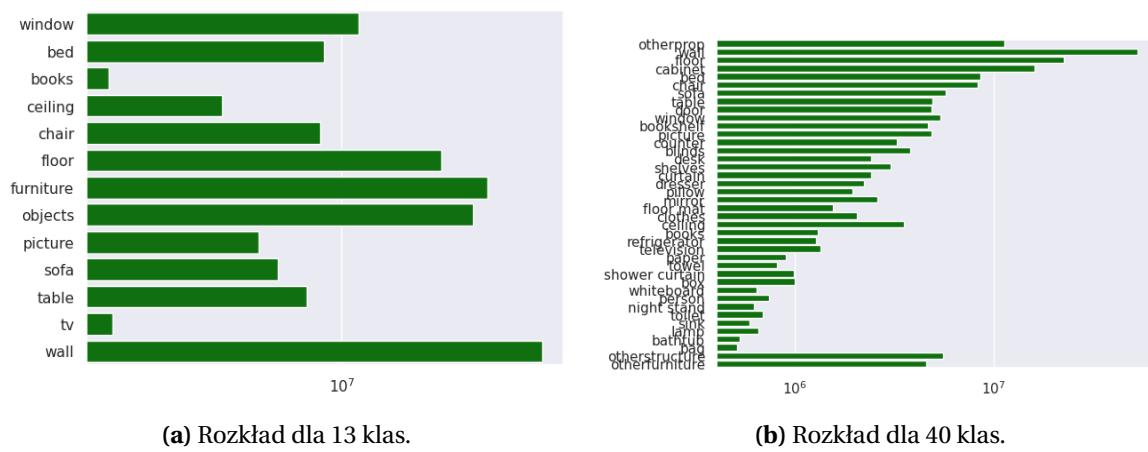
EDA przeprowadzone na zbiorze NYUv2 dostarczyło wielu interesujących szczegółów. W zbiorze domyślnie znajduje się 795 przykładów trenujących oraz 654 przykładów testujących. Ze zbioru testowego wyodrębniono zbiór walidacyjny stanowiący 20% zbioru testowego. Ponadto sprawdzono rozkład klas na przeszstrzeni 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 scalania 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).

4.3. Opis eksperymentów

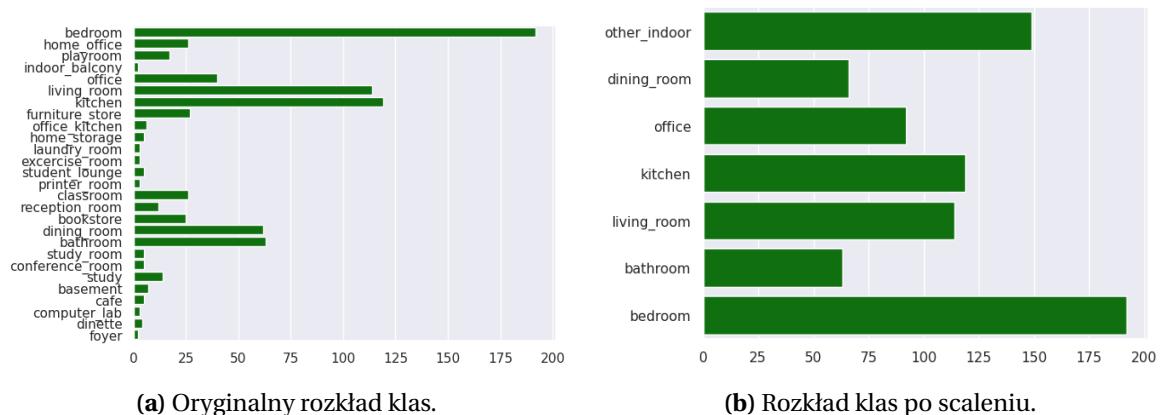
Przygotowanie danych

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, Ana-



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



Rysunek 4.3. Porównanie rozkładu klas dla zadania klasyfikacji sceny.

lysis and Application” Lei et. al. [13], autorzy udowadniają, że normalizacja stabilizuje i przyśpiesza trening oraz prawdopodobnie prowadzi do poprawy generalizacji.

Normalizacja jest ważnym krokiem przetwarzania wstępnego 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. Główka klasyfikacyjna przedstawia się jako sieć w pełni połączona (FC) z dwiema warstwami.

TO TRZEBA ZWIUZALIZOWAĆ!

Listing 1. Struktura głównej 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 pixeli.

Uczenie

ssssssssss

4.4. Wyniki

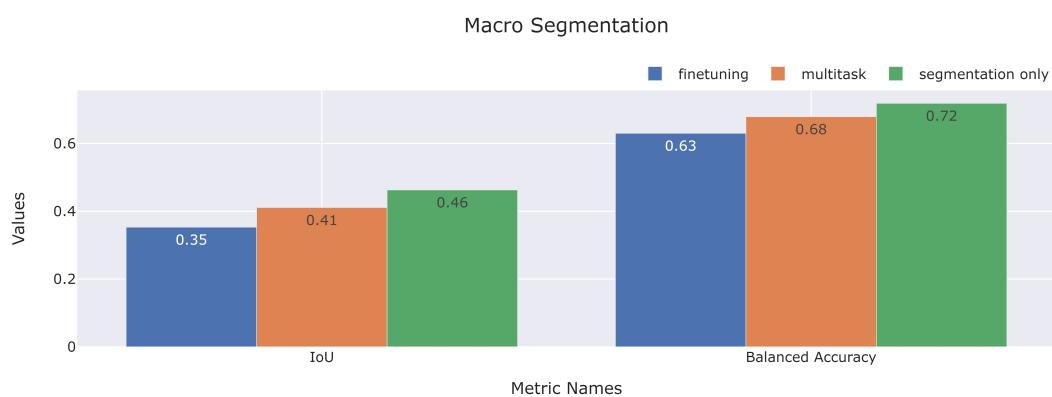
W tym rozdziale zostaną przedstawione empiryczne wyniki badań nad wspólną segmentacją semantyczną i klasyfikacją sceny w środowiskach wewnętrznych. Badania mają na celu opracowanie i ocenę różnych znanych i aktualnych technik uczenia głębokich sieci neuronowych. Aby to osiągnąć, przeprowadzono serię eksperymentów na zbiorze na reprezentacyjnych zbiorach danych. Analiza dotyczyła zarówno miar jakości sensu stricto jak i miar wydajnościowych proponowanych metod. Rozważono 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 zapewnią cenny wgląd w mocne strony i ograniczenia proponowanych metod.

4.4.1. Analiza miar jakości

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 powszechnie. Zatem makro analiza pokaże generalne rezultaty oraz otworzy dyskusję do dalszych, bardziej po głębiowych 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 odopowiedzieć 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.

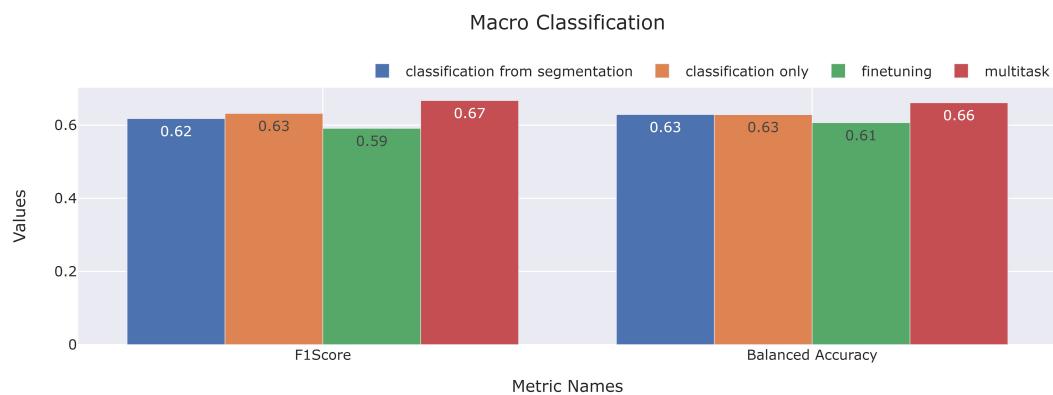


Rysunek 4.4. Porównanie miar IoU oraz dokładności dla segmentacji sceny.

Analizując rysunek 4.4 nie trudno zauważyc, że najlepsze rezultaty otrzymano w dla 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 miliar IoU jak i zbalansowanej dokładności (bAcc). Widać, że miliara IoU wypada gorzej niż bAcc. Wynik mogą sugerować, że trudno jest przeprowadzić transfer wiedzy z ImageNetu, gdyż finetunowanie wypada najsłabiej. Jest to naprawdopodobnie spowodowane zupełnieniem innym rozkładem klas dla wspomnianej bazie. Analiza sceny w przeciwnieństwie do klasyfikacji najczęściej cechuje się długogonowym rozkładem klas. Drugim istotnym szczegółem jest fakt, iż wagie dekodera i głowy segmentacyjnej są losowe. Uczenie wielozadaniowe zgodnie z zakładanymi wynikami nie polepsza segmentacji, gdyż łączna przeszczepienie segmentacji i klasyfikacji jest niewątpliwie trudniejsza do optymalizacji.

4. Eksperymenty

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

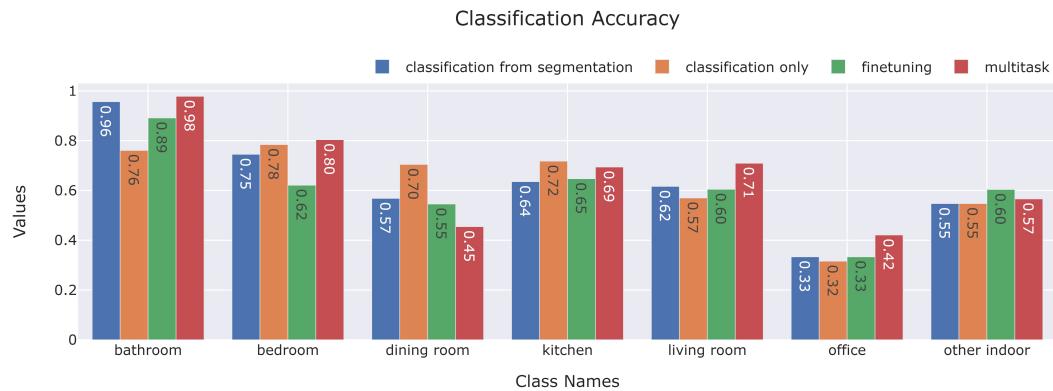


Rysunek 4.5. Porównanie miar F1 oraz dokładności dla klasyfikacji sceny.

Rezultaty przedstawia rysunek 4.5. Od razu da się zauważyć, że wyniki cechuje mniejsze odchylenie standardowe oraz, analizując łącznie miarę F1 oraz zbalansowaną dokładność, średnia. Fakt ten jest prawdopodobnie wynikiem znacznie mniejszej ilości parametrów uczących. Jako najlepszy rezultat uzyskuje uczenie wielozadaniowe. Ciekwy 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.

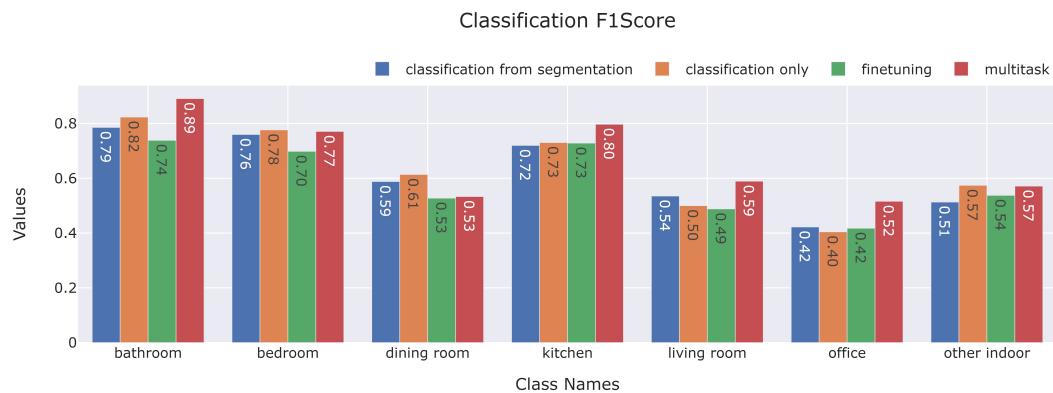
Analizowanie jakości algorytmu dla każdej z klas osobno jest ważne, ponieważ pozwala na bardziej szczegółowe zrozumienie mocnych i słabych stron algorytmu. Rozważając ogólną jakość algorytmu przy użyciu metryki makrośredniej, nie jest od razu jasne, w których klasach algorytm radzi sobie dobrze, a z 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 najgorsze 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 (rys.4.7) również nie jesteśmy w stanie wyróżnić faworyzowanej metody. W porównaniu



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

z wcześniejszej analizowaną dokładnością widać, że uczenie wielozadaniowe utrzymuje w większości przypadku bardzo dobre rezultaty. Widać też, że wyniki w obrębie każdej z klas mało różnią się między sobą.

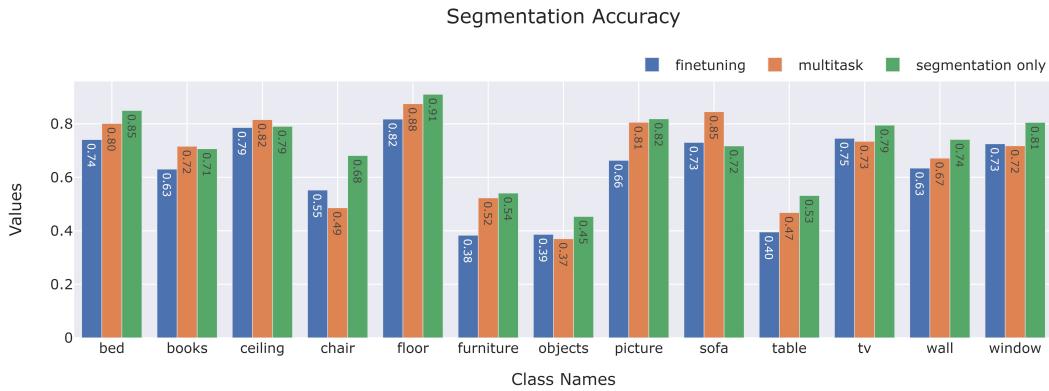


Rysunek 4.7. Porównanie miary F1 dla klasyfikacji sceny z rozróżnieniem konkretnych klas.

Analizując rysunek 4.8 przedstawiający 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 książka, sufit, sofa. Przypadek funetunowania nigdy nie osiągnął najlepszego rezultatu.

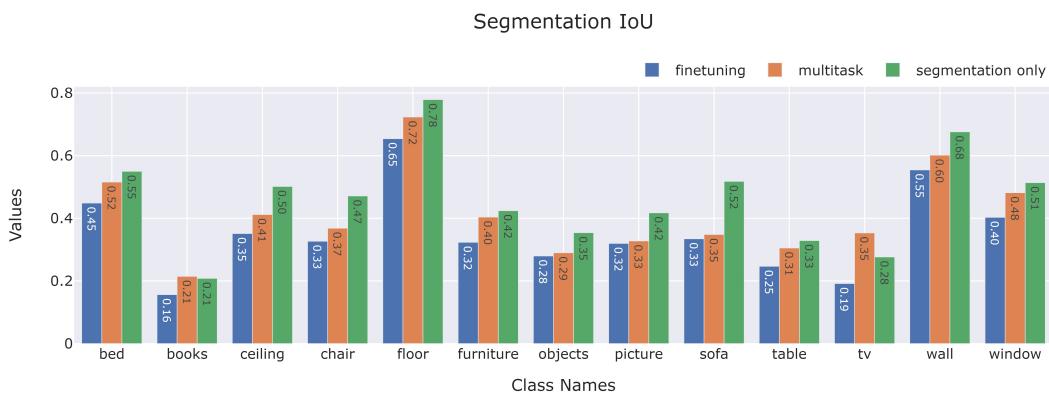
Na rysunku 4.9 przedstawiono IoU dla segmentacji semantycznej. Widać tutaj dużą dysproporcję między klasami podłoga, ściana, a pozostałymi klasami. Jest to zrozumiałe, klas te występują stosunkowo często na obrazie. Uczenie wyłącznie segmentacji uzyskuje

4. Eksperymenty



Rysunek 4.8. Porównanie dokładności segmentacji z rozróżnieniem konkretnych klas.

najlepsze wyniki na wszystkich klasach z wyłączeniem książek oraz telewizorów. W tych przypadkach najlepsze okazuje się uczenie wielozdaniowe.



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

4.4.2. Analiza czasowa

Ostatnio coraz częściej mówi się o zapotrzebowaniu na zasoby sprzętowe podczas uczenia maszynowego. Głębokie sieci neuronowe, a szczególnie te przetwarzające obrazy wymagają coraz więcej zasobów obliczeniowych do prawidłowego działania. Wynika to z dwóch głównych czynników. Po pierwsze duże modeli wizji komputerowej posiadają milionach parametrów. Drugim powodem jest przetwarzanie wielu obrazów, które de facto są zbiorem macierzy. Wiedzie to do większego zainteresowania zużywanymi zasobami podczas treningu oraz inferencji. W tym rozdziale przedstawię analizę czasu treningu oraz inferencji.

Analizując czas uczenia w przypadku kolejnych metod odkrywamy zalety transferu wiedzy oraz uczenie wielozadaniowe (tab. 4.1). Suma czasów uczenie wyłącznie seg-

mentacji oraz wyłącznie klasyfikacji (około 360s) znaczco przewyższa pozostałe metody. Najbardziej opłacalną czasowo metodą okazuje się finetuning. Jednakże na podstawie wyników miar jakości nie można uznać go za najbardziej optymalny. Pozostają jeszcze dwie metody - nauczenie segmentacji oraz dalsze uczenie klasyfikacji (ok. 260s) oraz uczenie wielozadaniowe (ok. 211s). Segmentacja a potem klasyfikacja osiąga najlepsze wyniki na segmentacji oraz przeciętne na klasyfikacji. Z drugiej strony uczenie wielozadaniowe osiąga najlepsze rezultaty na klasyfikacji oraz drugi najlepszy wynik na segmentacji. Łącząc to z faktem krótszego uczenia, można wysunąć wniosek, że uczenie wielozadaniowe jest optymalne pod względem czasu trwania oraz dawanych rezultatów.

nazwa zadania	czas[s]
tylko segmentacja	188.70
tylko klasyfikacja	170.47
klasyfikacja z segmentacją	70.78
finetuning	158.46
uczenie wielozadaniowe	210.97

Tabela 4.1. Porównanie czasu uczenia.

4.5. Porównanie jakości

Anliza metryk, czy różnych miar jakości jest niezbędna do ewaluacji zadań uczenia maszynowego. Odpowiedni wybór tych miar gwarantuje pełen informacji wgląd stanowiąc cenny wskazówki ewaluacyjne. Nie mniej nie wyklucza to istoty sprawdzenia rezultatów przez ludzkie oko. Mimo, że trudno byłoby przeglądać i ewaluować wiele zdjęć w dużych zbiorach danych, przekrojowe sprawdzenie jest kluczowe w analizie. Dostarcza bowiem wielu cennych, nieujętych w matematycznych formułach obserwacji. W tym rozdziale przedstawione zostaną rezultaty na wybranych zdjęciach.

4.5.1. Segmentacja semantyczna

Segmentacja semantyczna jest zadanie niewątpliwie trudnym. Jednocześnie równie ciężko jest określić dobrą funkcję jakości, uwzględniającą takie właściwości jak gładkość, dokładność czy precyzja segmentacji. Można oczywiście korzystać z wielu funkcji jakości, jednak ostateczny werdykt warto przejrzeć ręcznie. W połączeniu z wiedzą dotyczącą między innymi trudności klasyfikacji danej grupy pikseli lub niejednoznacznością niektórych grup pikseli po obejrzeniu nawet kilkunastu zdjęć, jesteśmy w stanie wykuć pewne wnioski.

Lazienka

Analizując rysunek 4.10 widzimy, że klasa przedmioty (ang. objects) jest bardzo szeroko rozumiana przez twórców zbioru danych. Wynika z tego fakt, że grupa ta nie posiada ścisłe określonych cech, które byłyby łatwo identyfikowalne. Model w tym przypadku połączył w sposób szeroki omawianą klasę. Ciekawą obserwacją jest zaznaczenie przez model

4. Eksperymenty

klasy krzesło. Po głębszej analizie można przypuszczać, że zlew ma podobną teksturę oraz kształt to metalowego krzesła. Klasy ściana, podłoga oraz meble została dość precyjnie sklasyfikowana.



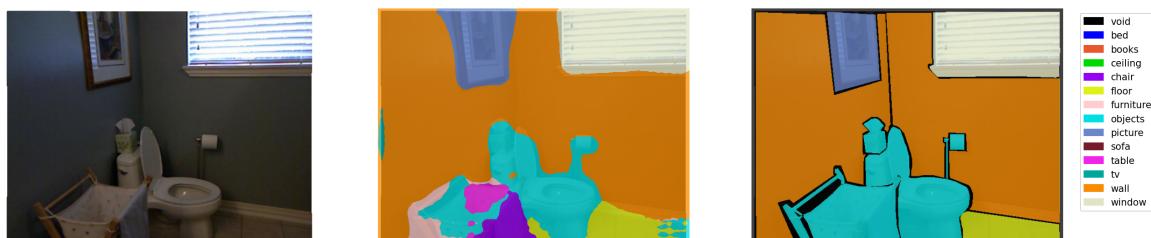
Rysunek 4.10. Porównanie jakości segmentacji dla klasy łazienka.

Sytuacja jest równie interesująca w przypadku rysunku 4.11. Model dopatruje się klasy meble w okolicach drzwi oraz przy zlewie. Pierwszy przypadek jest całkiem zrozumiały. Drewniane drzwi co do faktury mogą przypominać meble, na przykład drzwi od szafki. W drugiej sytuacji można domniemywać, że meble były często związane z umywalką czy nawet zlewem kuchennym, stąd model chętnie te klasy przydziela. Interesujące jest przydzielenie przez model etykiety obraz do włącznika światła.



Rysunek 4.11. Porównanie jakości segmentacji dla klasy łazienka.

Ostatnim obraz, przedstawiający łazienkę pokazuje rysunek 4.12. Tak jak wcześniej wspomniano ściany oraz podłoga są często dobrze klasyfikowane. Tak też jest w tym przypadku. Kosz na pranie okazał się być wyzwaniem. Model doszukiwał się tu takich obiektów jak stół, krzesło czy mebel.



Rysunek 4.12. Porównanie jakości segmentacji dla klasy łazienka.

Salon

Salon jest najczęściej reprezentowany przez duży pokój, w którym znajdują się kanapa, stolik z przedmiotami, krzesła/fotele oraz ściany z zawieszonymi obrazkami. Nie brakuje tutaj mebli i wielu obiektów.

Rysunek 4.13 jest przykładem częstego problemu adnotacji zdjęć. Często okazuje się, że dana grupa pikseli przedstawia więcej niż jedną klasę. Obraz oczekiwany przedstawia regał z książkami jako mebel. Model stwierdził jednak, wcale się nie myląc, że są to książki. Trudno się nie zgodzić z tą predykcją. Oznacza to, że zbiór jest poniekąd wewnętrznie sprzeczny w jakiejś części. Widać, że częstym kłopotem jest odróżnienie mebla od stołu. Zadowala fakt pierwszoplanowej kanapy, która bez poduszki została bardzo dobrze sklasyfikowana. Równie dobre rezultaty otrzymujemy dla klasy podłoga, sufit oraz obrazy. Dziwi natomiast fakt zaznaczenia fotela jako krzesła.



Rysunek 4.13. Porównanie jakości segmentacji dla klasy salon.

Na kolejnym rysunku 4.14 sytuacja jest nieco gorsza. Model miał trudności ze wskazaniem stolika na środku, który klasifykuje jako część kanapy. Sporo problemów wygenerowała klasa krzesło. Obrazy, ściany i podłoga zostały poprawnie sklasyfikowane. Przedmioty drugoplanowe, szczególnie dalsze, a więc mniejsze pozostały dla modelu jednakie.

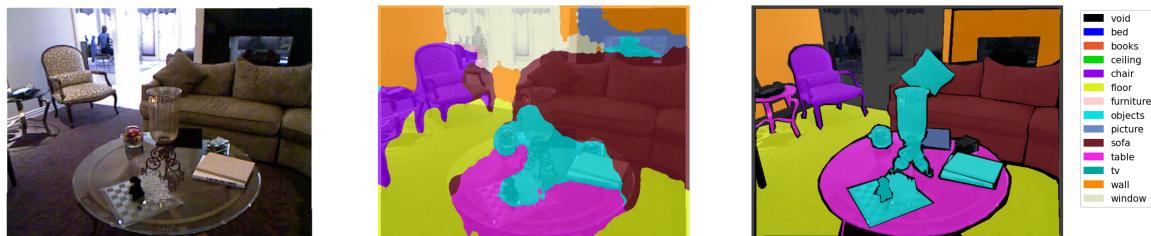


Rysunek 4.14. Porównanie jakości segmentacji dla klasy salon.

Scena salonu (rys. 4.15) jest znacznie lepiej sklasyfikowana niż poprzednia. Oprócz całkiem dobrze sklasyfikowanego stołu, obiektów i kanapy jest jeden ciekawy aspekt. Mianowicie obrazy docelowe znajdujące się w głębi obrazu zostały pominięte, czyli przedstawione jako pusta (ang. void). Mimo to klasyfikator celnie nadaje im klasy ściana oraz okna. To bardzo dobry prognostyk.

Sypialnia

4. Eksperymenty



Rysunek 4.15. Porównanie jakości segmentacji dla klasy salon.

Sypialnia to miejsce bardzo złożone. Jednak do charakterystycznych punktów tej sceny należą: łóżko, krzesło, stół oraz szafkę, które w przybliżeniu zostały całkiem dobrze sklasyfikowane. Brak zastrzeżeń budzą również klasę łóżko, podłoga oraz obiekty. Niewątpliwe ciekawe jest poprawne zaznaczenie okna, nawet w porze nocy. Jest to szczególnie cenna informacja, bo czarny prostokąt mógłby być zaklasyfikowany jako na przykład telewizor. Okno zostało zaznaczone zbyt szeroko, mianowicie fałszywie uznając prawdopodobnie lampa za okno. Prawdopodobnie kolor miał tu duże znaczenie.

na pierwszym planie rysunku 4.16 dwie krzesła, stół oraz szafkę, które w przybliżeniu zostały całkiem dobrze sklasyfikowane. Brak zastrzeżeń budzą również klasę łóżko, podłoga oraz obiekty. Niewątpliwe ciekawe jest poprawne zaznaczenie okna, nawet w porze nocy. Jest to szczególnie cenna informacja, bo czarny prostokąt mógłby być zaklasyfikowany jako na przykład telewizor. Okno zostało zaznaczone zbyt szeroko, mianowicie fałszywie uznając prawdopodobnie lampa za okno. Prawdopodobnie kolor miał tu duże znaczenie.



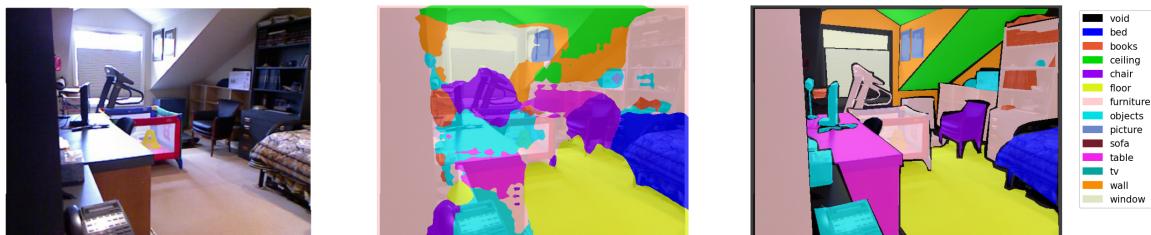
Rysunek 4.16. Porównanie jakości segmentacji dla klasy sypialnia.

Rysunek 4.17 to typowe zdjęcie sypialni. Czarną ramę łóżka model uznał za telewizor. Gdyby wyciąć samą tę ramę, wybór rzeczywiście nie byłby oczywisty. Poza tym łóżko zostało oznaczone całkiem poprawnie. Dziwi fakty sufitu w miejscu ściany, dla maski docelowej. Model słusznie przypisał tam ścianę. Obrazy zostały poprawnie oznaczone. Na zdjęciu widać, całkiem poprawną próbę klasyfikacji krzesła.



Rysunek 4.17. Porównanie jakości segmentacji dla klasy sypialnia.

Ostatni rysunek (rys. 4.18) był większym wyzwaniem dla modelu. Widać to szczególnie w przypadku pierwszoplanowego biurka. Model nie był w stanie podjąć decyzji co do ostatecznej klasy. Standardowo podłoga oraz sufit zostały sklasyfikowane prawidłowo. Nie inaczej było w przypadku klasy łóżko.



Rysunek 4.18. Porównanie jakości segmentacji dla klasy sypialnia.

Jadalnia

Obrazy związane z jadalnia to głównie sceny związane ze stołami oraz krzesłami.

Taką sytuację ma też miejsce na rysunku 4.19. Właściwie trudno tutaj znaleźć coś szczególnie interesującego. Cały obraz został całkiem dobrze pogrupowany. Wałpliwości budzi jedynie przypisanie do żyrandola klasy obrazy. Prawdopodobnie obrazy znajdujące się obok miały na to wpływ.



Rysunek 4.19. Porównanie jakości segmentacji dla klasy jadalnia.

Przypadek rysunku 4.20 wydaje się być ciekawszym. Szczególnie warte uwagi są tutaj okna na których znajdują się odbicia lustrzane. Refleksy są w wizji komputerowej zagadnieniem od dawna poruszany i znany. Można jednoznacznie stwierdzić, że trudno sobie poradzić w takich sytuacjach. Model prawdopodobnie mając trudności z tym obszarem przypisał go do klasy obiekt. Oprócz tego widzimy problemy z krzesłami w prawym dolnym rogu. Jasna, poleskująca skóra rzeczywiście przypomina nieco płytki podłogowe.

Ostatnim analizowanym obrazem w jadalni jest rysunek 4.21. Na pewno klasyfikacja klas takich jak stół, krzesła czy okno jest tutaj poprawna. Co więcej nie można tego do tego grona niezaliczyć klasy podłoga oraz sufit. Jedyny problem z grupowaniem na tym zdjęciu dotyczy samego roku zdjęcia, gdzie nie przyporządkowano klasy mebel. Pozostałe instancje tej klasy są poprawnie sklasyfikowane.

Kuchnia



Rysunek 4.20. Porównanie jakości segmentacji dla klasy jadalnia.



Rysunek 4.21. Porównanie jakości segmentacji dla klasy jadalnia.

Obrazy przedstawiające kuchnie to głównie zabudowa kuchni oraz sprzęt kuchenny. Czasen występuje tutaj na przykład stół z krzesłami.

Obraz 4.22 nie wydaje się trudnym do klasyfikacji, jednak pojawiło się tutaj kilka kwestii ważnych omówienia. Oprócz problemów z klasyfikacją stołu z prawej strony, który bardziej wygląda jak szafki z blatem w kuchni, obserwujemy błędne przypisanie tapety naściennej do klasy obrazy. Poza tym drewniane drzwi model klasyfikuje jako bardziej mebel niż ścianę, co ze względu na teksturę nie jest aż tak złym wyborem. Reszta zdjęcia została pogrupowana poprawnie.



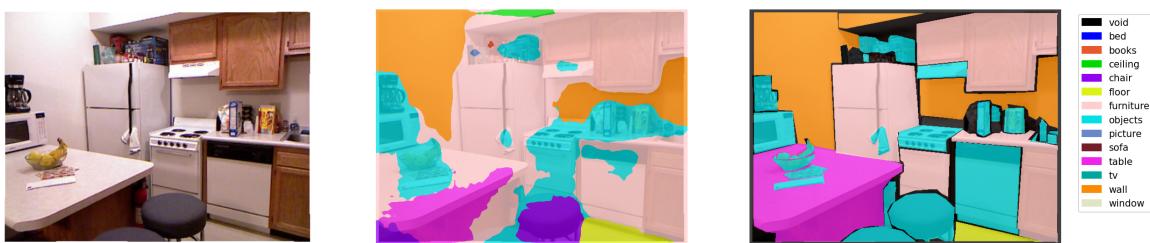
Rysunek 4.22. Porównanie jakości segmentacji dla klasy kuchnia.

Na rysunku 4.23 widzimy typową wąską kuchnię. Rezultaty są w miarę zadowalające poza przypisaniem lodówki do klasy obraz. Prawdopodobnie miały na to wpływ zdjęcia zawieszone na lodówce. Ściany, szafki i sufit zostały zaklasyfikowane prawidłowo.

Trzecim rysunkiem jest rys. 4.24. Największe wyzwanie stanowią tutaj obiekty zlokalizowane w różnych miejscach. Cieszy fakt, że mimo iż autorzy błędnie ocenili krzesło jako obiekt, model i tak zaznaczył go poprawnie. Widzimy tutaj również próbę klasyfikacji stołu. Powraca wtedy dyskusja na temat czy stół jest meblem tak jak został zresztą zaklasyfikowany.



Rysunek 4.23. Porównanie jakości segmentacji dla klasy kuchnia.



Rysunek 4.24. Porównanie jakości segmentacji dla klasy kuchnia.

Biuro

Sceny związane z biurem najczęściej przedstawiają biurka z krzesłami, zarówno w faktycznych biurach, o których często świadczy wykładzina, jak i w domowych pokojach typu biuro.

Na rysunku 4.25 widać scenę przedstawiającą pokój z drukarkami. Model dość dobrze radzi sobie ze ścianami oraz z podłogą, której akurat w tym przypadku nie ma zbyt wiele. Ciekawa jest wizja autorów zbioru danych określających mapę jako obiekt zamiast obrazu. Może trudno bez wahania przypisać wiszącej mapie miano obrazu, ale na pewno szybciej można ją określić jako plakat co można tłumazyć na angielski jako picture.

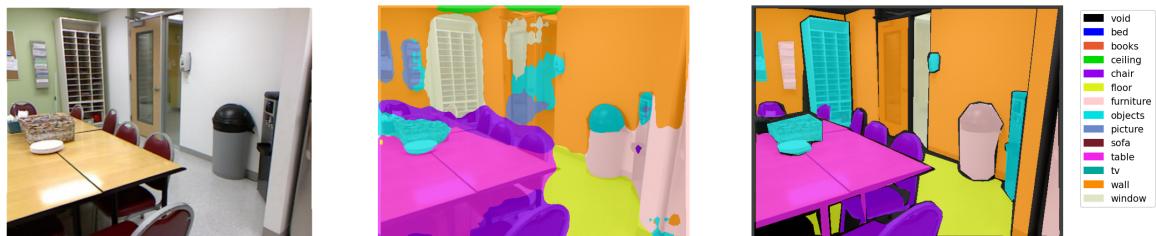


Rysunek 4.25. Porównanie jakości segmentacji dla klasy biuro.

Rysunek 4.26 przedstawia salę konferencyjną. Lewa strona obrazu została zaklasyfikowana całkiem poprawnie. Wyzwaniem dla modelu okazał się prawy dolny róg, gdzie należało przypisać klasę kolejno od lewej mebel, obiekt, ściana, co model uprościł do prostego mebla. To zdecydowanie zła klasyfikacja.

Ostatnim rysunkiem 4.27 jest pomieszczenie przedstawiające najprawdopodobniej biuro domowe. Klasyfikacja okan, podłogi oraz ściany była prawie bezbłędna. Gorzej

4. Eksperymenty



Rysunek 4.26. Porównanie jakości segmentacji dla klasy biuro.

model porawdził sobie z stołem, który po części sklasyfikował jako telewizor ze względu na bardzo ciemny oraz prostokąty charakter.

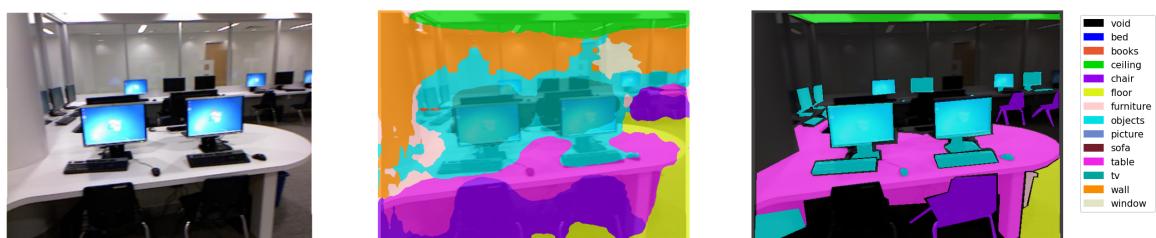


Rysunek 4.27. Porównanie jakości segmentacji dla klasy biuro.

Inne pomieszczenia

Sceny związane z klasą inne pomieszczenia budzą najwięcej wątpliwości. Nie wiadomo bowiem, co dokładnie może się tam znaleźć.

Na rysunku 4.28 znajduje się wspólna przestrzeń biurowa. Widzimy, że znajduje się tutaj wiele obszarów typu void, zatem model dokładnie nie wie co powinno się tam znaleźć. Dziwi to szczególnie w przypadku pierwszego krzesła po prawej stronie. Nie mniej jednak model dość dobrze zgaduję tę klasę. Jest zrozumiałym, że pokój otoczony ścianami. W gruncie rzeczy szklana szyba oczywiście jest ścianą w tym przypadku. Model niezbyt dobrze pogrupował klasę obiekty. Jest tutaj wiele do poprawy.



Rysunek 4.28. Porównanie jakości segmentacji dla klasy inne pomieszczenia.

Rysunek 4.29 przedstawia pomieszczenie biurowe. Wszystkie obrazki zostały zaklasyfikowane poprawnie. Okna zostały przypisane jako obrazy. Model dobrze pogrupował człowieka. Wyzwanie stanowiła klasa obiekty.



Rysunek 4.29. Porównanie jakości segmentacji dla klasy inne pomieszczenia.

Ostatnim analizowanym rysunkiem segmentacji jest 4.30. Meble oraz obrazy, jak również podłoga zostały poprawnie zaklasyfikowane. Całkiem dobra jest segmentacja krzesła na pierwszym planie. Problemy wystąpiły z drugim krzesłem oraz częścią kanapy. Model błędnie doszukiwał się klasy stół na tym obrazie.



Rysunek 4.30. Porównanie jakości segmentacji dla klasy inne pomieszczenia.

4.5.2. Klasyfikacja sceny

Podobnie jak w przypadku segmentacji semantycznej trudno jest czasem określić jednoznacznie jakość modelu bazując wyłącznie na miarach jakości. W nieniejszym rozdziale zostaną przytoczone wszystkie błędne klasyfikacje z podziałem na konkretne klasy. Pozwoli to wysnuć pewne obserwacje na temat podobieńst tych klas oraz pomoże wysunąć wnioski co do tych błędów. Co więcej przedstawione zostaną statystyki błędnej klasyfikacji, aby lepiej zobrazować te błędy.

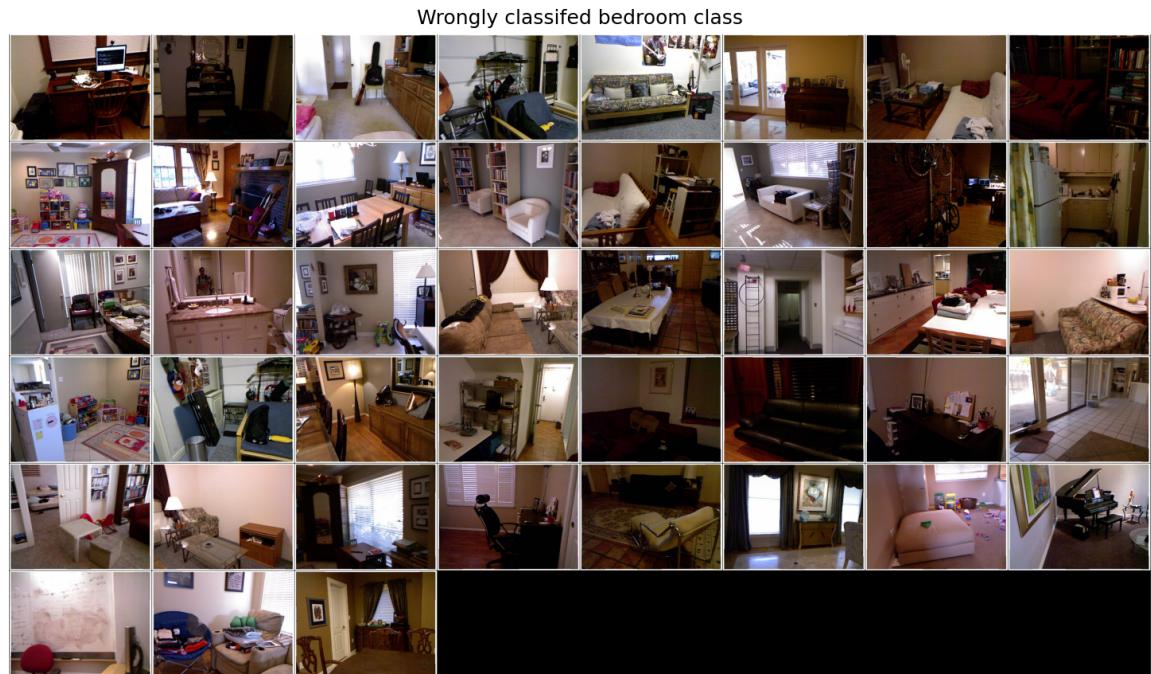
Na rysunku 4.31 przedstawiono 10 błędnych przypisań dla klasy łazienka. Dziewięć z dziesięciu błędów dotyczyło klasy kuchnia. Można doszukiwać się, że kuchnia jak i łazienka ma poniekąd podobny schemat. Na pewno występuje te same klasy jak zlew czy meble. Tylko raz klasyfikator uznał że sypialnia jest łazienką. Najwięcej pomyłek algorytm popełnił na klasie sypialnia (rys.4.32). Na pewno wynika to z faktu, iż była to klasa dominująca. Co drugi błąd następował na klasie sypialnia. Szczególnie często gdy łóżkiem była kanapa lub na zdjęciu występował fotel. Ponad 32% błędów w sumie stanowiły klasy biuro oraz jadalnia. Można przypuszczać, że tym razem kluczowym elementem świadczącym o predykcji był stół. Jadalnia była pomylona w sumie 11 razy (rys. 4.33). Zgodnie z przedstawionym rysunkiem, na większości zdjęć występuje stół i krzesła.

Najmniej pomyłek jest dla klasy kuchnia (rys. 4.34). Klasy inne pomieszczenia oraz

4. Eksperymenty



Rysunek 4.31. Porównanie jakości klasyfikacji dla klasy łazienka.



Rysunek 4.32. Porównanie jakości klasyfikacji dla klasy sypialnia.



Rysunek 4.33. Porównanie jakości klasyfikacji dla klasy jadalnia.

kuchnie zostały błędnie zaklasyfikowane jako biuro w większości przypadków. Trudno określić, skąd akurat wynikają takie rezultaty.

Model najczęściej błędnie przypisywał klasę inne pomieszczenia dla biura w mniej niż połowie przypadków. Pozostałe przypadki należą do klas sypialnia oraz jadalnia. Klasa inne pomieszczenia jest szczególnie narażona na pomyłki, gdyż to połączenie najróżniejszych klas scen.



Rysunek 4.34. Porównanie jakości klasyfikacji dla klasy kuchnia.



Rysunek 4.35. Porównanie jakości klasyfikacji dla klasy biuro.



Rysunek 4.36. Porównanie jakości klasyfikacji dla klasy inne pomieszczenia.

Analiza błędnie sklasyfikowanej scen dostarczyła wielu ważnych informacji. Najczęściej przyczyną błędów było znaczne podobieństwo występujących klas przedmiotów między różnymi klasami scen.

5. Podsumowanie

-mozna bylo lepiej to zrobic: - augmentacja - więcej epok - regularyzacja smoothingu - regularazacja weight decay -zeby w pelni moc ocenic potencjal uczenia wielozadaniowe nalezialoby spawdzic modele na wielu zbiorach, w razie potrzeby zachowujac przestrzen reprezentacji a dosziflowujac ostatnie warstwy decyzyjbedroombedroom

niepoprawen klasy szum podczas uczenia klasa obiekt generuja problemy klasy szero-kie generuja problemy Analiza przewidywań dostarczyła wielu cennych szczegółów, które byłoby trudno zauważyć patrząc jedynie na liczby. drewno to meble obrazy zawsze były na powierzchniach takich jak sciany czy szafki REZULTATAT słabe wyniki bo - sprzeczność klas (regał mebel czy książki, czy mebel i stół) - klasa objects

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] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik i S. Savarese, “Which tasks should be learned together in multi-task learning?”, w *International Conference on Machine Learning*, PMLR, 2020, s. 9120–9132.
- [9] J. Yao, S. Fidler i R. Urtasun, “Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation”, w *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, s. 702–709.
- [10] 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.
- [11] 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.
- [12] D. Seichter, S. B. Fischedick, M. Köhler i H.-M. Groß, “Efficient Multi-Task RGB-D Scene Analysis for Indoor Environments”, w *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, s. 1–10. DOI: 10.1109/IJCNN55064.2022.9892852.

5. Bibliografia

- [13] 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 wewnętrz klasowej oraz wieloznaczności semantycznej [4].	15
2.2 Segmentacja wewnętrz pomieszczeń [5].	16
3.1 Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation (2012) [9].	19
3.2 Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification 2016 [10].	20
3.3 Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images 2018 [11].	21
3.4 Efficient Multi-Task RGB-D Scene Analysis for Indoor Environments [12] . . .	21
3.5 Architektura wielozadaniowej sieci.	22
3.6 Arhitekrura sieci szeregowej.	24
4.4 Porównanie miar Iou oraz dokładnosci dla segmentacji sceny.	29
4.5 Porównanie miar F1 oraz dokładnosci dla klasyfikacji sceny.	30
4.6 Porównanie dokładności klasyfikacji sceny z rozróżnieniem konkretnych klas. . .	31
4.7 Porównanie miary F1 dla klasyfikacji sceny z rozróżnieniem konkretnych klas. .	31
4.8 Porównanie dokładności segmentacji z rozróżnieniem konkretnych klas. . . .	32
4.9 Porównanie miary IoU segmentacji z rozróżnieniem konkretnych klas.	32
4.10 Porównanie jakości segmentacji dla klasy łazienka.	34
4.11 Porównanie jakości segmentacji dla klasy łazienka.	34
4.12 Porównanie jakości segmentacji dla klasy łazienka.	34
4.13 Porównanie jakości segmentacji dla klasy salon.	35
4.14 Porównanie jakości segmentacji dla klasy salon.	35
4.15 Porównanie jakości segmentacji dla klasy salon.	36
4.16 Porównanie jakości segmentacji dla klasy sypialnia.	36
4.17 Porównanie jakości segmentacji dla klasy sypialnia.	36
4.18 Porównanie jakości segmentacji dla klasy sypialnia.	37
4.19 Porównanie jakości segmentacji dla klasy jadalnia.	37
4.20 Porównanie jakości segmentacji dla klasy jadalnia.	38
4.21 Porównanie jakości segmentacji dla klasy jadalnia.	38
4.22 Porównanie jakości segmentacji dla klasy kuchnia.	38
4.23 Porównanie jakości segmentacji dla klasy kuchnia.	39
4.24 Porównanie jakości segmentacji dla klasy kuchnia.	39
4.25 Porównanie jakości segmentacji dla klasy biuro.	39
4.26 Porównanie jakości segmentacji dla klasy biuro.	40
4.27 Porównanie jakości segmentacji dla klasy biuro.	40
4.28 Porównanie jakości segmentacji dla klasy inne pomieszczenia.	40

4.29 Porównanie jakości segmentacji dla klasy inne pomieszczenia.	41
4.30 Porównanie jakości segmentacji dla klasy inne pomieszczenia.	41
4.31 Porównanie jakości klasyfikacji dla klasy łazienka.	42
4.32 Porównanie jakości klasyfikacji dla klasy sypialnia.	42
4.33 Porównanie jakości klasyfikacji dla klasy jadalnia.	42
4.34 Porównanie jakości klasyfikacji dla klasy kuchnia.	43
4.35 Porównanie jakości klasyfikacji dla klasy biuro.	43
4.36 Porównanie jakości klasyfikacji dla klasy inne pomieszczenia.	43

Spis tabel

4.1 Porównanie czasu uczenia.	33
---------------------------------------	----