

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 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	10
2.1. Klasyfikacja sceny	10
2.2. Segmentacja obrazu	10
2.3. Głębokie uczenie i konwolucje	11
2.4. Uczenie wielozadaniowe	11
3. Rozwiązanie	12
3.1. Przegląd rozwiązań	12
3.2. Zarys rozwiązania problemu	13
4. Eksperymenty	14
4.1. Opis zbioru danych	14
4.2. Analiza zbioru danych	15
4.3. Opis eksperymentów	16
4.4. Wyniki	17
5. Podsumowanie	19
Bibliografia	21
Spis rysunków	22
Spis tabel	22

1. Wprowadzenie

1.1. Cel pracy

Uzyskanie informacji o środowisku wewnętrz budynków poprzez:

- klasyfikację pomieszczenia
- segmentację semantyczną

1.2. Motywacje

Istnieje wiele powodów, dla których temat pracy jest wart uwagi.

Po pierwsze rozwiązanie może być wykorzystane w nawigacji robota. Wykrywanie przeszkód jest kluczowym aspektem możliwości poruszania się robota. Zostanie ono podjęte przez zadanie segmentacji. Należy zwrócić uwagę, że robot powinien zachowywać się ostrożniej w kuchni oraz w łazience. Ta informacja zostanie uzyskana poprzez klasyfikację sceny.

Innym zastosowanie rozważanego rozwiązania jest pomoc dla osób niewidomych. Osoba niepełnosprawna mogłaby wówczas poruszać się po środowisku domowym z większą łatwością, mając na sobie kamerę oraz informację o otaczającej przestrzeni.

2. Wstęp teoretyczny

2.1. Klasifikacja sceny



Rysunek 2.1. Problem różnorodności wewnętrzklasowej oraz wieloznaczności semantycznej [1].

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-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.

Zadanie klasyfikacji sceny jest trudne ze względu na 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.

2.2. Segmentacja obrazu

Zadanie segmentacji obrazu to przyporządkowanie każdemu pikselowi etykietę (rys. 2.2). W rezultacie obraz zostaje podzielony na homogeniczne regiony pod względem pewnych własności. Zadanie segmentacji można rozszerzyć do zadania segmentacji instancji (ang. instance segmentation), czyli segmentacji klasycznej rozszerzonej o rozróżnienie 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 jednakowa. Segmentacja instancji pozwala natomiast takie rozróżnienie uczynić.



Rysunek 2.2. Segmentacja wewnętrz pomieszczeń [2].

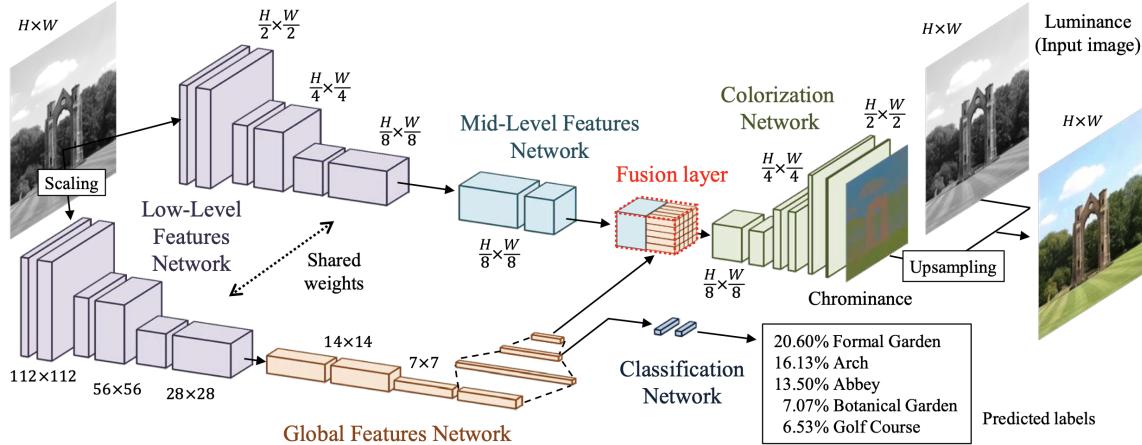
Segmentacja semantyczna w dalszej części pracy będzie odnosić się do klasycznej wersji. Segmentacja instancji nie jest tematem pracy.

2.3. Głębokie uczenie i konwolucje

2.4. Uczenie wielozadaniowe

3. Rozwiązanie

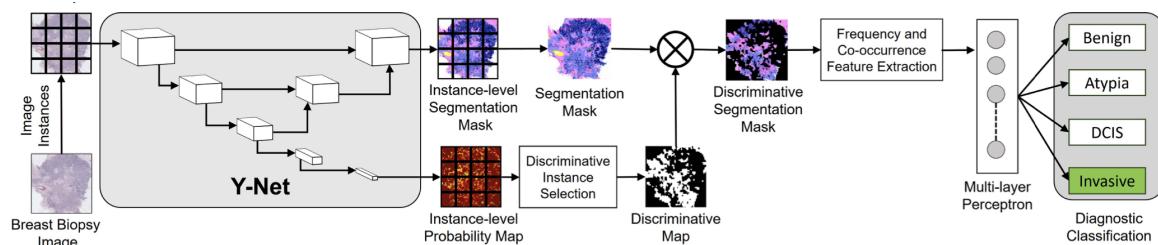
3.1. Przegląd rozwiązań



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 [3].

Współcześnie do zadań wizji komputerowej używa się głębkich 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 [3]” 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 [4].

Kolejnym artykułem jest „Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images 2018 [4]”. 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.

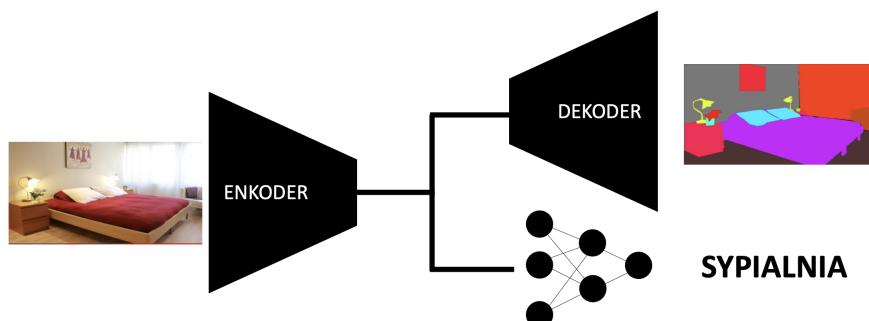
3.2. Zarys rozwiązania problemu

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.

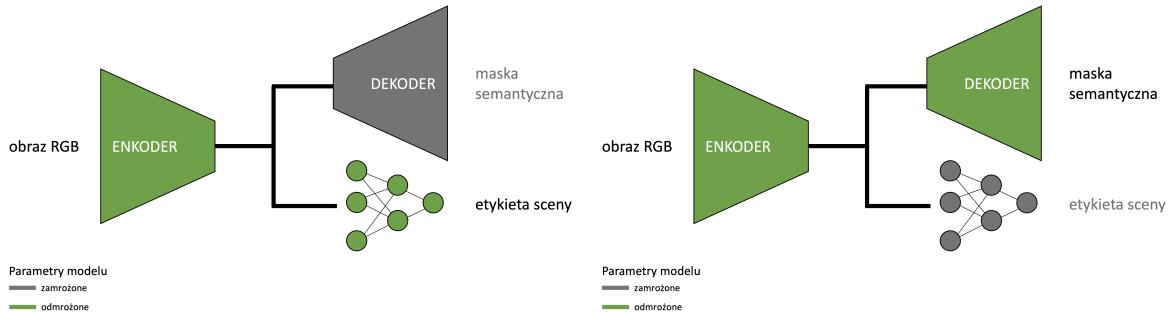
Mając na uwadze, że symultaniczne uczenie może negatywnie wpływać na jakość uczenia obu zadań, eksperymenty przeprowadzono etapowo. Pierwszym etapem było uczenie jednozadaniowe. Eksperymenty polegały na sprawdzeniu jakości segmentacji oraz klasyfikacji osobno. Wykorzystano do tego tą samą architekturę, 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 wagi aktualizowane były zgodnie z zagregowaną stratą.

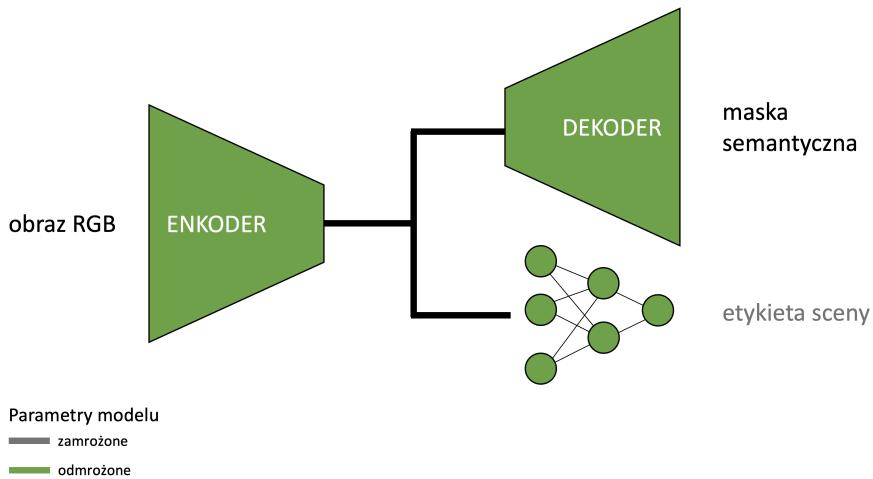
Ostatecznie porównano jakość na przesztreningu obu etapów.



Rysunek 3.3. Architektura sieci zastosowana w pracy inżynierskiej.



Rysunek 3.4. Podejście jednozadaniowe.

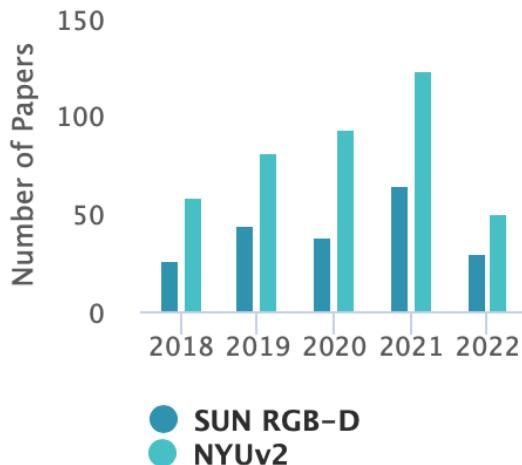


4. Eksperymenty

4.1. Opis zbioru danych

Zbiór danych powinien ściśle odpowiadać założeniom postawionym w pracy. Inferencja wymaga użycia kamery Kinect. Zatem zbiór danych powinien zawierać kategorie scen, segmentacje obrazów oraz najlepiej być ujętym przez kamerę Kinect wersji pierwszej.

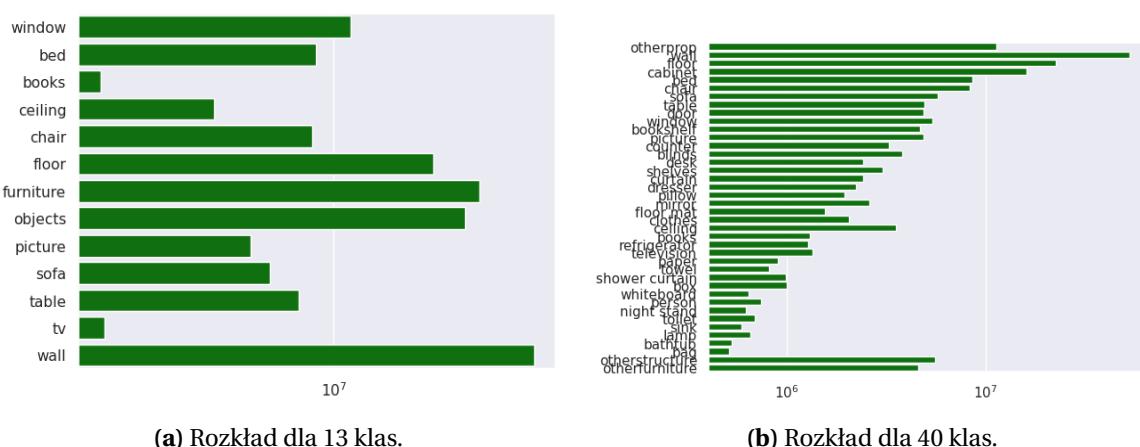
Po prześledzeniu wielu zbiorów danych udało się sprostać powyższym wymaganiom, uzyskując dwa podobne zbiory danych - NYUv2 oraz SUN RGBD. Ostatecznie wybrano NYUv2 z uwagi, że zbiór ten został zawierać zdjęcia pomieszczeń, w których nie są posprzątane. Fakt ten uznano, za ważny, iż uważano, że będzie 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]

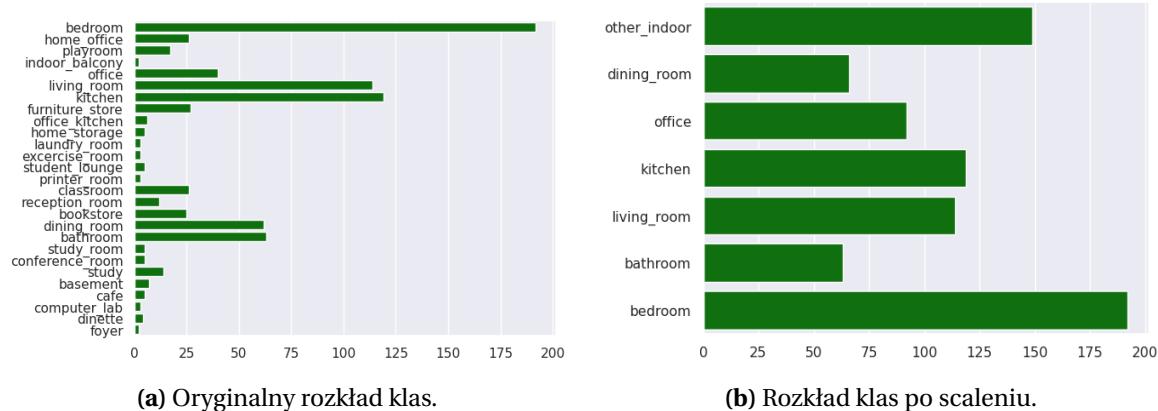
4.2. Analiza zbioru danych

Przeprowadzono eksploracyjną analizę danych dla obu zadań. W zbiorze znajduje się 795 przykładów trenujących oraz 654 przykładów testujących. 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 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).



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

4. Eksperymenty



(a) Oryginalny rozkład klas.

(b) Rozkład klas po scaleniu.

Rysunek 4.3. Porównanie rozkładu klas dla zadania klasyfikacji sceny.

4.3. Opis eksperymentów

Przygotowanie danych

Obrazy RGB zostały poddane jedynie normalizacji ze średnią (0.485, 0.456, 0.406) oraz odchyleniem standardowym (0.229, 0.224, 0.225).

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),  
10    nn.Softmax(dim=1),
```

Funkcja straty

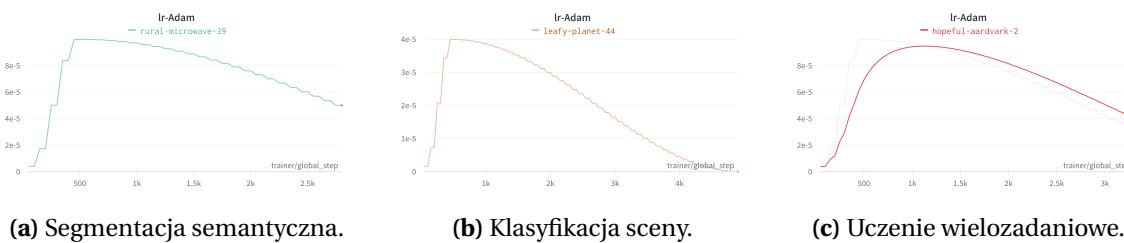
W obu przypadkach jako funkcję straty wykorzystano ważoną entropię skrośną. Wagi odzwierciedlały odwrotność liczności w zbiorze. Dla klasyfikacji liczona była ilość klas, natomiast dla segmentacji ilość pixeli.

Uczenie

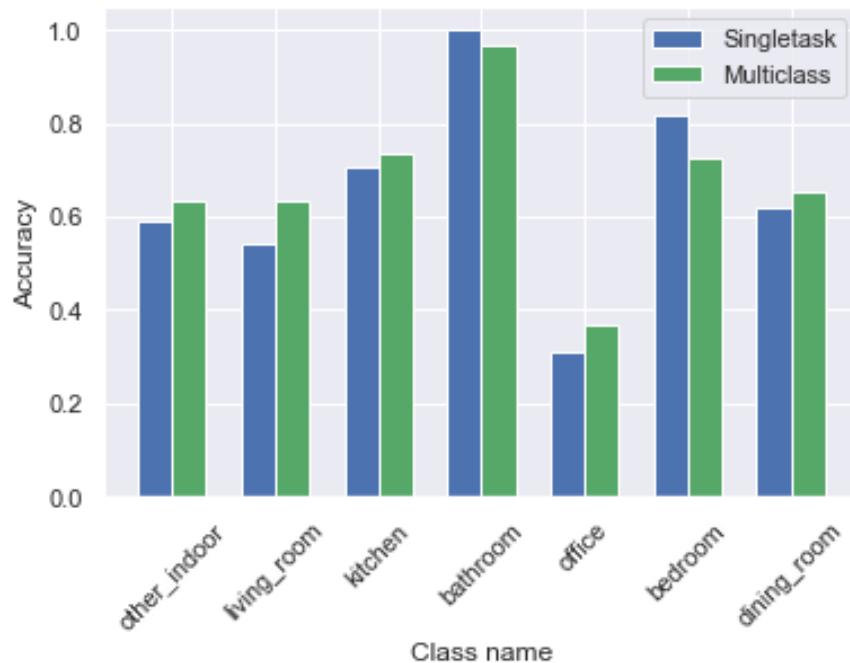
zadanie/[%]	Acc jednozadaniowe	Acc wielozadaniowe
segmentacja	67.87	67.48 -0.39
klasyfikacja	65.50	67.45 +1.95
średnia	66.69	67.47 +0.78

Tabela 4.1. Porównanie dokładności dla uczenia jedno- i wielozadaniowe.

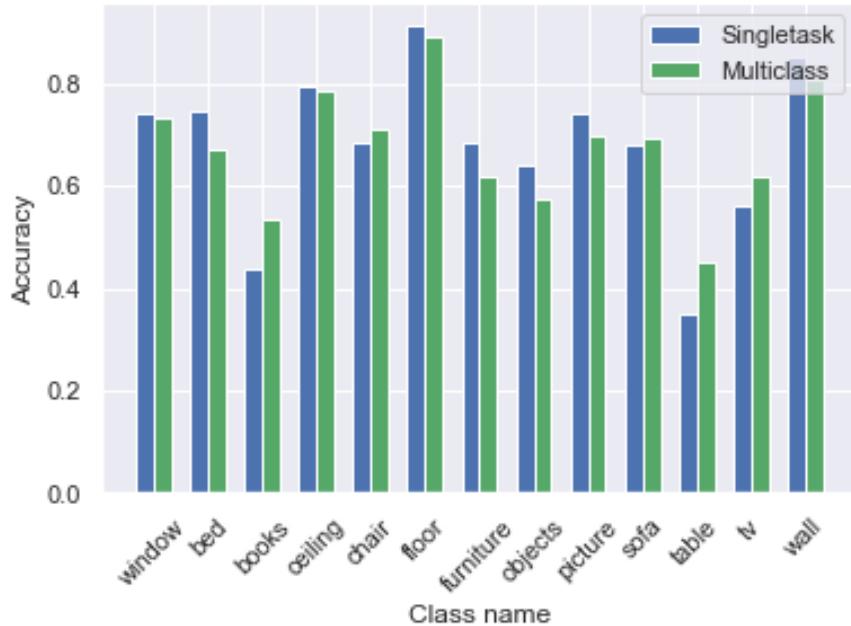
Uczenie odbywało się co najwyżej 50 epok aż do ustalenia się straty na zbiorze walidacyjnym. Krok uczenia był zmienny zgodnie z polityką One Cycle (rys.4.4).

**Rysunek 4.4.** One cycle learning rate scheduler policy

4.4. Wyniki

**Rysunek 4.5.** Porównanie dokładności dla każdej z klas w zadaniu klasyfikacji pomieszczeń

Uczenie wielozadaniowe w rozważanym przypadku nieznacznie poprawia wyniki sieci (tab. 4.1). Dla zadania segmentacji semantycznej otrzymujemy spadek jakości o 0.39



Rysunek 4.6. Porównanie dokładności dla każdej z klas w zadaniu segmentacji semantycznej

punkta procentowego. Zadanie klasyfikacji poprawia się o 1.95 p.p. w porównaniu z uczeniem jednozadaniowym. Ostatecznie otrzymujemy zysk na poziomie 0.78 punktu procentego na średniej z zadań. Poprawa jest niewielka, jednak jest to duży sukces biorąc pod uwagę, że mamy do dyspozycji 2 razy mniej parametrów niż w przypadku dwóch osobnych sieci. Przekłada się to bezpośrednio na czas inferencji, który w przypadku robotyki i systemów czasu rzeczywistego jest kluczowy.

Wartym zobaczenia jest fakt, iż uczenie wielozadaniowe poprawia wyniki dla klas które osiągają najsłabsze rezultaty w uczeniu jednozadaniowym. Poprawie ulega klasa office (rys. 4.5) dla klasyfikacji oraz klasy books oraz table (rys. 4.6) dla segmentacji semantycznej. Powodem jest prawdopodobnie mniejsze obciążenie (bias) modelu spowodowane faktem wzajemnej regularyzacji obu zadań w procesie uczenia. Innymi słowy, model ma mniejszą tendencję do przeuczenia.

5. Podsumowanie

Bibliografia

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.

Spis rysunków

2.1 Problem różnorodności wewnątrzklasowej oraz wieloznaczności semantycznej [1].	10
2.2 Segmentacja wewnątrz pomieszczeń [2].	11
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 [3].	12
3.2 Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images 2018 [4].	12
3.3 Architektura sieci zastosowana w pracy inżynierskiej.	13
3.5 Architektura sieci jako uczenie wielozadaniowego.	14
4.5 Porównanie dokładności dla każdej z klas w zadaniu klasyfikacji pomieszczeń	17
4.6 Porównanie dokładności dla każdej z klas w zadaniu segmentacji semantycznej	18

Spis tabel

4.1 Porównanie dokładności dla uczenia jedno- i wielozadaniowe.	17
-------------------------------------------------------------------------	----