



Imię i nazwisko studenta: Bartosz Bieliński  
Nr albumu: 165430  
Studia pierwszego stopnia  
Forma studiów: stacjonarne  
Kierunek studiów: Automatyka i Robotyka  
Profil: Systemy decyzyjne i robotyka

Imię i nazwisko studenta: Piotr Winkler  
Nr albumu: 165504  
Studia pierwszego stopnia  
Forma studiów: stacjonarne  
Kierunek studiów: Automatyka i Robotyka  
Profil: Systemy decyzyjne i robotyka

## **PRACA DYPLOMOWA INŻYNIERSKA**

Tytuł pracy w języku polskim: Zastosowanie sieci neuronowych do edycji obrazów

Tytuł pracy w języku angielskim: Application of neural networks for image editing

Potwierdzenie przyjęcia pracy	
Opiekun pracy	Kierownik Katedry/Zakładu (pozostawić właściwe)
<i>podpis</i>	<i>podpis</i>
dr inż. Mariusz Domżalski	

Data oddania pracy do dziekanatu:

**POLITECHNIKA GDAŃSKA**

Katedra Systemów Decyzyjnych i Robotyki

**PRACA INŻYNIERSKA**

**Zastosowanie sieci neuronowych do edycji  
obrazów**

**Autorzy**

Piotr Winkler

Bartosz Bieliński

Gdańsk 2019

## **STRESZCZENIE**

Tematem pracy jest zbadanie możliwości zastosowania sieci neuronowych do edycji obrazów. Głównym celem było stworzenie narzędzi opartych na wyuczonych sieciach neuronowych, służących do odpowiedniego przetwarzania i modyfikowania obrazu. <Zakres pracy >

<Zastosowane metody badań ><wyniki ><najważniejsze wnioski >

**Słowa kluczowe:** sieć neuronowa, przetwarzanie obrazu, konwolucyjna sieć neuronowa, splotowa sieć neuronowa, model generatywny, głęboka sieć neuronowa,

**Dziedzina nauki i techniki zgodna z OECD:** Nauki inżynierijne i techniczne, Elektrotechnika, elektronika, inżynieria informatyczna, Sprzęt komputerowy i architektura komputerów

## **ABSTRACT**

The subject of the work is research on the possibilities of applying neural networks for image editing. The main goal was to create tools based on trained neural networks used for appropriate processing and modifying of images. <Zakres pracy >

<Zastosowane metody badań ><wyniki ><najważniejsze wnioski >

**Keywords:** neural network, image processing, convolutional neural network, generative adversarial network, deep neural network

**Field of science and technology in accordance with the requirements of the OECD:** Engineering and technology, Electrical engineering, Electronic engineering, Information engineering, Computer hardware and architecture

## WYKAZ WAŻNIEJSZYCH OZNACZEŃ I SKRÓTÓW

- ANN (ang. Artificial Neural Network) - Sztuczna sieć neuronowa
- DNN (ang. Deep Neural Network) - Głęboka sieć neuronowa
- FCL (ang. Fully Connected Layer) - Warstwa gęstwa
- CNN (ang. Convolutional Neural Network) - Splotowa sieć neuronowa
- FCN (ang. Fully Convolutional Network) - Sieć w pełni splotowa
- GAN (ang. Generative Adversarial Network) - Sieci o modelu generatywnym
- VAE (ang. Variational Autoencoder) - Autoenkodery wariacyjne
- ReLU (ang. Rectified Linear Unit) - Jednostronnie obcięta funkcja liniowa
- BatchNorm (ang. Batch Normalization) - Normalizacja zbioru danych pogrupowanych w pakiety
- YUV - Model barw, w którym Y odpowiada za luminancję obrazu, a UV są to dwa kanały chrominacji i kodują barwę
- IcGAN (ang. Invertible conditional Generative Adversarial Network) - Odwracalny warunkowy model generatywny
- cGAN (ang. conditional Generative Adversarial Network) - warunkowy model generatywny
- Dropout (!!! ang. spadkowicz - tłumaczył Piotr Winkler) - technika regularyzacji mająca na celu ograniczać przeuczanie się sieci neuronowych przez zapobieganie złożonej koadaptacji danych treningowych
- PRelu (ang. Parametric Rectified Linear Unit) - Parametryczna jednostronnie obcięta funkcja liniowa
- RReLU (ang. Randomized Leaky Rectified Linear Unit) - Losowo nieszczelna jednostronnie obcięta funkcja liniowa
- Adam (ang. Adaptive Moment Estimation) - Adaptacyjna estymacja pędu.
- SGD (ang. Stochastic Gradient Descent) - Metoda stochastycznego spadku gradientowego.
- AdaGrad (ang. Adaptive Gradient Algorithm) - Adaptacyjny algorytm gradientowy.

# Spis treści

<b>WYKAZ WAŻNIEJSZYCH OZNACZEŃ I SKRÓTÓW</b>	<b>4</b>
<b>1 Wstęp i cel pracy</b>	<b>6</b>
1.1 Cel pracy . . . . .	6
1.2 Dotychczasowe dokonania . . . . .	6
1.3 Założenia projektowe . . . . .	7
1.4 Układ pracy . . . . .	7
<b>2 Podstawy teoretyczne</b>	<b>8</b>
2.1 Sieci splotowe . . . . .	8
2.2 FCN . . . . .	10
2.3 Modele generatywne . . . . .	10
2.4 Autoenkodery . . . . .	11
<b>3 Przegląd rozwiązań</b>	<b>12</b>
3.1 Colorful image colorization . . . . .	12
3.2 Image Style Transfer Using Convolutional Neural Networks . . . . .	13
3.3 Invertible Conditional GANs for image editing . . . . .	14
3.4 Neural photo editing . . . . .	15
<b>4 Zaimplementowane rozwiązania</b>	<b>17</b>
4.1 Automatyczne kolorowanie czarno-białych obrazów . . . . .	18
4.1.1 Podejście . . . . .	18
4.1.2 Model podstawowy . . . . .	19
4.1.3 BatchNorm . . . . .	20
4.1.4 Dropout . . . . .	21
4.1.5 Modyfikacja rozdzielczości . . . . .	22
4.1.6 Wykorzystywany zbiór treningowy . . . . .	23
4.1.7 Przetwarzanie wstępne danych . . . . .	23
4.1.8 Przetwarzanie końcowe danych . . . . .	24
4.1.9 Augmentacja danych . . . . .	25
4.1.10 Funkcje kosztów . . . . .	26
4.1.11 Funkcje aktywacji . . . . .	28
4.1.12 Algorytmy optymalizacyjne . . . . .	28
4.1.13 Trening . . . . .	29
4.1.14 Rezultaty . . . . .	29
<b>5 Podsumowanie</b>	<b>31</b>
<b>Bibliografia</b>	<b>32</b>
<b>Spis rysunków</b>	<b>33</b>
<b>Spis tabel</b>	<b>33</b>

# 1 WSTĘP I CEL PRACY

Sztuczne sieci neuronowe sięgają swym początkiem lat 40. XX wieku. Historia ich rozwoju odnotowała trzy okresy, w których rozwiązania te odbijały się szerokim echem w środowisku naukowym.

Pierwszy model neuronu, a potem perceptron zapoczątkowały rozwój tej dziedziny nauki, jednak pierwsze sieci jednowarstwowe nie były w stanie rozwiązywać złożonych problemów. Przeszkodę nie do pokonania stanowiła dla nich nawet prosta funkcja logiczna XOR. Z tego powodu badania sieci neuronowych zostały na długi czas porzucone.

Pojawienie się algorytmu wstępnej propagacji błędów pozwalającego skutecznie uczyć wielowarstwowe sieci neuronowe ponownie wzmożło zainteresowanie tematem, jednak tym razem na drodze postępowi stanęły ograniczenia technologiczne ówczesnych czasów.

Wreszcie wraz z nadaniemiem XXI wieku postępujący rozwój komputerów oraz internetu umożliwił sztucznym sieciom neuronowym rozwinięcie skrzydeł. Wejście w erę "big data" otworzyło dostęp do olbrzymich zbiorów danych niezbędnych do treningu sieci, a pojawienie się wysokowydajnych jednostek obliczeniowych pozwoliło znacznie ten proces przyspieszyć.

Zapoczątkowany w ten sposób rozwój trwa do dnia dzisiejszego. Sztuczne sieci neuronowe odnajdują zastosowanie w wielu dziedzinach życia i nauki. Grają w gry, przeprowadzają symulacje, przewidują i prognozują zachowania rynku, czy pogody, analizują i przetwarzają obrazy cyfrowe.

Z punktu widzenia niniejszej pracy największe znaczenie ma oczywiście ostatni z wymienionych punktów. Zdefiniowanie sieci neuronowych, jako matematycznych modeli obliczeniowych ujawnia ich naturalne predyspozycje do pracy na obrazach cyfrowych. W praktyce stanowią one bowiem zbiór liczb, wartości poszczególnych pikseli, który sieć neuronowa jest w stanie analizować, przetwarzać i modyfikować.

## 1.1 Cel pracy

Celem pracy jest stworzenie szeregu narzędzi programistycznych oferujących szeroki wachlarz możliwości edytowania obrazu. Narzędzia te oparte mają być na technologii sieci neuronowych. W szczególności przetestowana będzie skuteczność rozwiązań dedykowanych do przetwarzania obrazów, takich jak sieci konwolucyjne albo modele generatywne.

Po opracowaniu tychże narzędzi, opisane zostaną efekty pracy oraz zbadana zostanie skuteczność sieci neuronowych jako rozwiązania nakreślonej problematyki. Omówione zostaną także wykorzystane architektury zaimplementowanych modeli, wykorzystane funkcje kosztu, metody aktualizowania wag sieci oraz przebiegi treningu modeli.

## 1.2 Dotychczasowe dokonania

<Syntetyczny opis dotychczasowych dokonań w danej tematyce? >

### ***1.3 Założenia projektowe***

Główym założeniem pracy było zaprojektowanie i zaimplementowanie narzędzi programistycznych służących do edycji obrazu. Narzędzia te muszą wykorzystywać do swoich celów nauuczone sieci neuronowe. Na podstawie jakości działania tychże narzędzi, oceniona zostanie ich rzetelność oraz skuteczność.

Następnie przeprowadzona zostanie analiza słuszności zastosowania sieci neuronowych jako rozwiązania przedstawionej problematyki. Uzyskane rozwiązanie zostanie także zestawione ze znanyymi algorytmami do edycji obrazu nie opierającymi się na technologii sieci neuronowych.

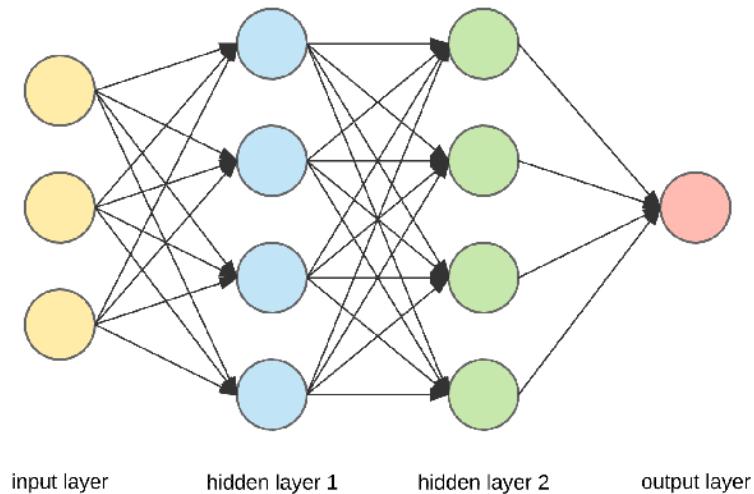
### ***1.4 Układ pracy***

W pierwszym rozdziale przedstawiono zarys rozwoju Sieci Neuronowych na przestrzeni lat oraz niezbędne podstawy teoretyczne związane z przedstawioną problematyką i wybranym dla niej rozwiązaniem.

W kolejnym rozdziale dokonano przeglądu już istniejących rozwiązań, opisano ich przeznaczenie, sposób działania oraz uzyskane rezultaty. Oceniono także wpływ danego rozwiązania na rozwój sieci neuronowych w dziedzinie przetwarzania i edytowania obrazów.

## 2 PODSTAWY TEORETYCZNE

W dzisiejszych czasach sieci neuronowe zajmują ważną pozycję na rynku narzędzi do edycji obrazu. Jest to głównie spowodowane ich umiejętnością do reprodukowania i modelowania niesłiniowych procesów, a także nowoczesnymi technikami przetwarzania plików graficznych. Jednak pierwsze architektury ANN (ang. artificial neural network) nie nadawały się do przetwarzania grafik. Było to częściowo spowodowane faktem, że obrazy, będące w rzeczywistości macierzami wartości pikseli, ciężko było skutecznie podać na wejście typowych architektur DNN (ang. deep neural network) zbudowanych pierwotnie z wielu warstw ukrytych, pomiędzy którymi połączenia są na zasadzie każdy z każdym oraz mają swoje wagi podlegające modyfikacji w trakcie procesu uczenia. Taka struktura pokazana została na Rysunku 2.1. Obrazy o niskiej rozdzielcości można było przekształcić w wektory wartości poszczególnych pikseli i w takiej postaci podawać na wejście sieci, jednak w przypadku obrazów o wyższej rozdzielcości to rozwiązanie, ze względu na znaczną długość powstających wektorów, nie oferowało dobrych rezultatów. Dopiero nowe architektury sieci spowodowały przełom w tej dziedzinie. Wprowadzenie do najistotniejszych i najciekawszych z nich zostanie przedstawione w poniższym rozdziale, a także rozwinięte w dalszej części tej pracy.



Rysunek 2.1: Struktura DNN

### 2.1 Sieci splotowe

Neuronowe sieci splotowe (CNN ang. convolutional neural network) stanowią podstawową strukturę w zakresie przetwarzania i analizowania obrazów cyfrowych. Są to sieci o hierarchicznej strukturze stanowiące podwaliny większości klasyfikatorów, detektorów, czy sieci segmentujących.

Autorzy jednego z artykułów traktujących o sieciach splotowych [3] opisują je następująco:

*'CNN to skuteczny algorytm poznawczy, stosowany powszechnie przy rozpoznawaniu wzorców i przetwarzaniu obrazów. Posiada wiele cech, takich jak prosta struktura, mniej parametrów treningowych, czy zdolność do adaptacji. CNN stały się gorącym tematem w zakresie analizy głosu i rozpoznawania obrazu. Ich struktura oparta na podziale wag czyni je bardziej podobnymi do biologicznych sieci neuronowych. Redukuje to złożoność modelu sieci oraz liczbę wag'.*

Na CNN składają się zazwyczaj trzy rodzaje warstw, z których każda posiada inne cechy.

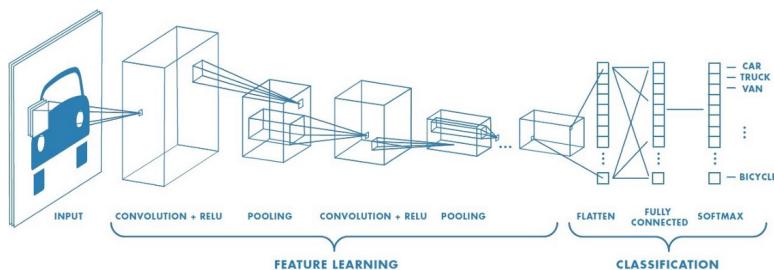
Podstawową warstwę stanowi warstwa splotowa. Składa się ona ze zbioru filtrów (neuronów) odpowiedzialnych za ekstrakcję cech z analizowanych obrazów poprzez dokonanie operacji konwolucji na obrazie poprzez przesuwanie zestawu filtrów wzdłuż niego. Na wyjściu filtrów otrzymuje się macierze o mniejszej rozdzielcości reprezentujące wyniki operacji konwolucji w danym punkcie. Każda kolejna warstwa splotowa wydobywa z obrazu cechy o wyższych poziomach abstrakcji bazując na wynikach obliczeń poprzednich warstw tego rodzaju. Dzięki temu procesowi kolejne warstwy filtrów uczą się rozpoznawać kluczowe cechy na obrazie, od drobnych elementów takich jak krawędzie albo kształty po bardziej złożone takie jak części ciała albo całe obiekty. Filtry te są zazwyczaj inicjowane losowymi wartościami i w miarę trenowania, dopasowują swoje parametry do wybranej problematyki.

Drugim istotnym elementem sieci splotowych jest warstwa poolingu. Może zostać opisana następująco [4]:

*'We wszystkich przypadkach pooling pomaga uczynić reprezentację w przybliżeniu niezmiennej w stosunku do małych tłumaczeń danych wejściowych. Niezmienność wobec tłumaczenia oznacza, że jeśli poddamy dane wejściowe nieznacznej translacji, to wartość większości wyników poddanych poolingu nie ulegnie zmianie'.*

Końcowy element CNN w większości przypadków stanowią warstwy gęste (FCL ang. Fully Connected Layer). Odpowiadają one za dokonanie odpowiedniej klasyfikacji obrazu na podstawie danych dostarczonych przez warstwy poprzedzające. Są przez to nieodzowne w przypadku zadań związanych z wszelkiego rodzaju klasyfikacją obrazów.

Wymienione tutaj elementy składowe sieci splotowych mogą przyjmować różne rozmiary i występować w różnych konfiguracjach, co przedstawiono na poniższym Rysunku 2.2.

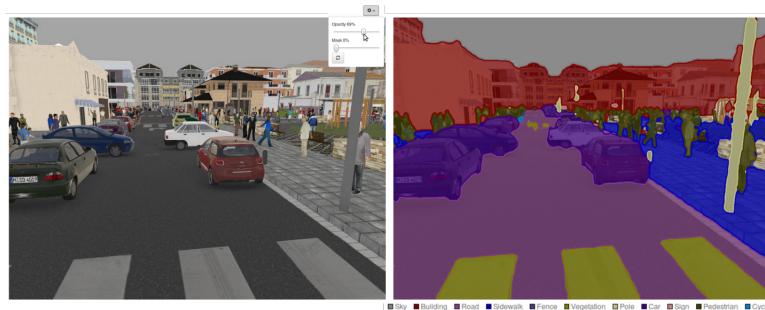


Rysunek 2.2: Przykładowa struktura CNN

Zapewnia to szerokie pole do eksperymentów i sprawia, że sieci te zdolne są rozwiązywać złożone, różnorodne problemy z wielu dziedzin codziennego życia.

## 2.2 FCN

Jednym z kluczowych problemów, jakie stawia przed badaczami edycja obrazów jest zagadnienie segmentacji semantycznej. Klasyczna klasyfikacja, polegająca na przypisywaniu obrazów do odpowiednich grup tematycznych, jest w tym przypadku sprowadzana do poziomu pojedynczych pikseli. Oznacza to, że sieci neuronowe przeznaczone do tego zadania są w stanie dokonać klasyfikacji dla każdego pojedynczego piksela analizowanego obrazu. Na tej podstawie uzyskiwany jest podział na segmenty, z których każdy reprezentuje inną klasę obiektów, jak na poniższym Rysunku 2.3.



Rysunek 2.3: Segmentacja semantyczna

W większości modele odpowiadające za przeprowadzanie segmentacji składają się z szeregowego połączenia enkodera oraz dekodera. Enkoder jest zazwyczaj pre-trenowaną siecią neuronową przeznaczoną do klasyfikowania obrazów. Dekoder odpowiada za semantyczne rzutowanie cech w niskiej rozdzielcości, wyuczonych przez enkoder, na wysoką rozdzielcość samych pikseli tworząc wspomniany wcześniej podział segmentowy.

FCN (ang. Fully Convolutional Networks) stanowią szczególny rodzaj sieci neuronowych przeznaczonych do segmentacji obrazów. Składają się one wyłącznie z kombinacji warstw splotowych oraz poolingu. Są w stanie przetwarzać obrazy o dowolnej, zmiennej wielkości, w odróżnieniu od innych typów modeli, w których zastosowanie warstw gęstych (FCL) wymusza z góry ustalone rozmiary danych wejściowych.

Naprzemienne przepuszczanie obrazów przez wspomniane warstwy splotowe oraz pooling może powodować niską rozdzielcość wyjściowych rezultatów pracy tych sieci oraz rozmycie granic poszczególnych obiektów. Z tego powodu w nowoczesnych rozwiązaniach stosuje się dodatkowe mechanizmy zapobiegające tego typu trendom.

## 2.3 Modele generatywne

Koncepcja modeli generatywnych, w skrócie GANów, przedstawiona została w 2014 roku przez Iana Goodfellow oraz jego współpracowników na uniwersytecie w Montrealu [1]. Modele te stanowią połączenie dwóch głębokich sieci neuronowych działających przeciwstawnie do siebie nawzajem.

Pierwsza sieć to tak zwany generator. W odniesieniu do tematu pracy, jego działanie polega na generowaniu nowych obrazów, lub ich fragmentów na podstawie wektora szumów.

Obrazy te przekazywane są, równolegle z zestawem obrazów prawdziwych, do dyskryminatora stanowiącego drugą część modelu GAN. Działanie tej sieci neuronowej polega na określeniu (w skali 0 do 1), w jakim stopniu produkty wyjściowe generatora odpowiadają obrazom rzeczywistym.

W opisany modelu występuje zatem podwójna pętla sprzężenia zwrotnego. Dyskryminator określa autentyczność obrazów porównując je ze zdefiniowaną odgórnie bazą danych. Z kolei generator otrzymuje informację o skuteczności swojego działania ze strony dyskryminatora.

Model generatywny znajduje się w stanie ciągłego konfliktu. Generator dąży do jak najdokładniejszego fałszowania obrazów w celu oszukania dyskryminatora, którego celem jest z kolei jak najdokładniejsze wykrywanie podróbek. Obie sieci neuronowe nieustannie dążą do osiągnięcia przewagi nad rywalem w procesie treningu. Ciągła rywalizacja sprawia, że zarówno generator, jak i dyskryminator zyskują coraz wyższą skuteczność działania.

W praktyce modele generatywne są w stanie naśladować dowolną dystrybucję danych. Są w stanie kreować światy podobne do naszego w zakresie obrazu, dźwięku czy mowy. Można powiedzieć, że są to prawdziwi syntetyczni artyści.

#### **2.4 Autoenkodery**

### 3 PRZEGŁĄD ROZWIĄZAŃ

Na przestrzeni ostatnich paru lat pojawiło się wiele rozwiązań zastosowania sieci neuronowych do edycji obrazu, duża część z nich była przełomowa w swojej dziedzinie. Powstawały rewolucyjne architektury sieci oraz technologie z nimi związane. Takie cechy sieci jak niezwykła zdolność do generalizacji zdobytej wiedzy na nowe przypadki oraz olbrzymia elastyczność sprawiły, że znalazły one wiele rzeczywistych zastosowań.

W tym rozdziale skupiono się na przedstawieniu kilku interesujących rozwiązań dla omawianej problematyki.

#### 3.1 *Colorful image colorization*

Wraz z rozwojem sieci neuronowych, rosło zainteresowanie możliwościami zastosowania ich do kolorowania czarno-białych obrazów. Jedno z dostępnych rozwiązań tego zagadnienia zostało przedstawione przez grupę pracowników Uniwersytetu w Berkeley [5]. Zamiarem ich pracy było stworzenie modelu, który niekoniecznie odtwarza oryginalne barwy obrazu, ale generuje barwy prawdopodobne, zdolne przekonać ludzkiego obserwatora o autentyczności obrazu. Uzyskane rezultaty zostały przedstawione na Rysunku 3.1.



Rysunek 3.1: Efekt kolorowanie czarno-białych zdjęć przez wytrenowany model.

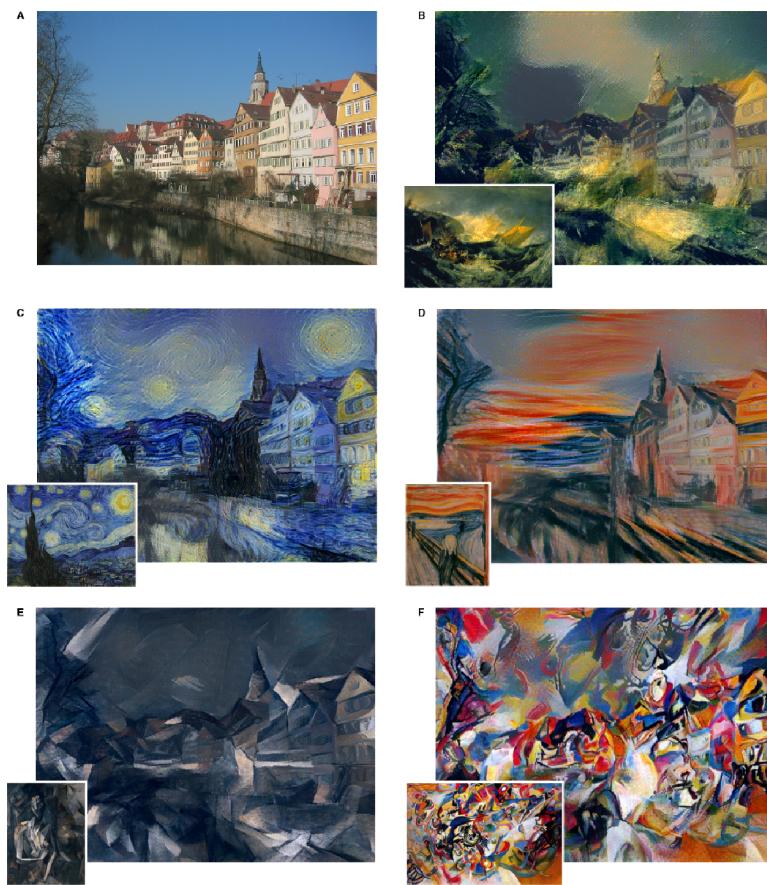
Wykorzystany model składa się z wielu warstw CNN, w których skład wchodzą warstwa filtrów konwolucyjnych, warstwa ReLU (ang. Rectified Linear Unit) oraz warstwa BatchNorm (ang. Batch normalization). Aby zapobiec utracie informacji przestrzennych, sieć nie posiada warstw łączących. Istotny był także sposób przygotowania zbioru danych do trenowania modelu. Obrazy ze zbioru uczącego były wpierw konwertowane do modelu YUV, a następnie kanał Y był podawany na wejście modelu, warstwy UV pełniły funkcję pożąданiej odpowiedzi w uczeniu nadzorowanym.

Ważnym aspektem zbadanym w artykule było także dobranie odpowiedniej funkcji kosztu. Nieuodpowiedni wybór skutkował desaturacją kolorowanych obrazów, jedną z potencjalnych przyczyn tego zjawiska może być tendencja sieci do tworzenia bardziej konserwatywnych odpowiedzi. Aby zniwelować ten efekt w modelu została zastosowana specjalna technika modyfikacji funkcji kosztu. Polega ona na przewidywaniu dystrybucji możliwych kolorów dla każdego piksela i zmienianiu kosztu dla modelu, w celu wyróżnienia rzadko spotykanych kolorów.

Powstałe rozwiązańe dowodzi olbrzymiego potencjału zastosowanie sieci neuronowych w dziedzinie pracy nad obrazami.

### 3.2 Image Style Transfer Using Convolutional Neural Networks

W roku 2016 został przedstawiony światu A Neural Algorithm of Artistic Style [6]. Wprowadził on przełom w dziedzinie przenoszenia stylu jednego obrazu na inny, a jego sukces opierał się na właściwym wykorzystaniu konwolucyjnych sieci neuronowych. Podstawą tego sukcesu było odkrycie przez Leona A. Gatys oraz jego współpracowników, że w CNN reprezentacja treści obrazu oraz jego stylu jest rozłączna. Umożliwia to wydobycie stylu przetwarzanego obrazu oraz połączenie go z treścią innego obrazu, czego dokonuje właśnie A Neural Algorithm of Artistic Style. Rezultaty takich operacji można zaobserwować na Rysunku 3.2



Rysunek 3.2: Obrazy będące kombinacją treści zdjęcia ze stylami kilku znanych dzieł sztuki.

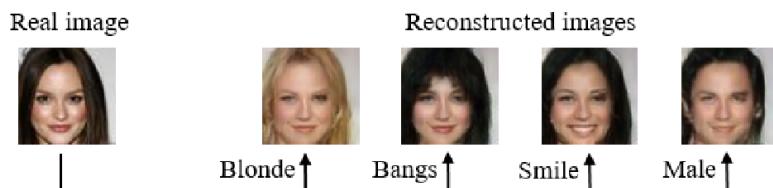
Do zbudowania modelu zostały użyte warstwy konwolucyjne oraz łączące z architektury VGG-Network [7], która została wytrenowana pod kątem rozpoznawanie obiektów i określania ich położenia. Dzięki temu sieć przetwarzając obraz tworzy jego reprezentację, która wraz z kolejnymi warstwami, przedstawia coraz wyraźniejszą informację o obiektach, a niekoniecznie o dokładnym wyglądzie obrazu. W modelu nie została użyta ani jedna warstwa gęsta, dzięki czemu na wyjściu możliwe jest otrzymanie dwuwymiarowego obrazu. Dla lepszej syntezy obrazów, w warstwach łączących zastosowano próbkowanie wartością średnią zamiast maksymalną. Takie zabiegi umożliwiają wyliczenie reprezentacji stylu z korelacji pomiędzy różnymi cechami w różnych warstwach konwolucyjnych.

Cały proces renderowania polega na odpowiednim zapisaniu w modelu treści oraz stylu obrazów otrzymanych przez wcześniejsze podanie na model tychże obrazów. Wpierw obraz, z którego pobierany jest styl, jest podawany na wejście sieci oraz przeliczany, reprezentacja stylu wyselekcjonowana z właściwych warstw jest przechowywana, obraz z treścią jest poddawany temu samemu procesowi, ale reprezentacja treści jest wyciągana z ostatnich warstw konwolucyjnych. W celu uzyskania fuzji obrazów, zapisane reprezentacje treści i stylu są zapisywane w tych warstwach modelu skąd zostały odczytane, a następnie na wejście podawany jest obraz składający się z losowego szumu białego. Następnie poprzez iteracyjną minimalizację funkcji kosztu, obraz wejściowy jest modyfikowany, co w rezultacie końcowym doprowadza do nałożenia zapisanego stylu na wczytaną treść.

A Neural Algorithm of Artistic Style jest świetnym przykładem, jak elastyczne mogą być interfejsy do modyfikacji obrazu oparte na technologii sieci neuronowych.

### 3.3 Invertible Conditional GANs for image editing

Edycja obrazów może być dokonywana na wielu różnych poziomach zaawansowania i abstrakcji, operacje takie jak nakładanie filtrów mogą być wykonywane przez proste algorytmy. Jednak w przypadku próby modyfikacji elementów na obrazie, algorytmy te nie będą w stanie dokonać semantycznych zmian ze względu na brak możliwości zrozumienia treści obrazu. Rozwiązanie tego problemu zostało przedstawione w postaci modelu IcGAN (ang. Invertible Conditional Generative Adversarial Network) w roku 2016 [8]. Zaprezentowany model był to enkoder z możliwością generowania wektora informacji o atrybutach obrazu połączony z warunkowym GAN zdolnym do kontrolowania cech generowanych obrazów na podstawie dodatkowej informacji warunkowej. Takie działanie umożliwia wprowadzania zmian w atrybutach generowanego obrazu uzyskiwanego na wyjście cGAN (ang. conditional Generative Adversarial Network). Rezultaty działania modelu można zaobserwować na Rysunku 3.3.



Rysunek 3.3: Obrazy generowane przez IcGAN.

Wykorzystany w IcGAN Ekonder w rzeczywistości składa się z dwóch podrzędnych Enkoderów, Enkoder  $E_z$  koduje wejściowy obraz do utajonego wektora  $z$  reprezentacji obrazu, natomiast Enkoder  $E_y$  generuje wektor informacji  $y$  oddających pewne kluczowe atrybuty obrazu. Enkody są trenowane z użyciem już wytrenowanego cGAN oraz obrazów rzeczywistych z etykietami ze zbioru uczącego. Zbadane zostały także różne podejścia interakcji między dwoma Enkoderami, wyróżnić można podejście, w którym Enkdodery są w pełni niezależne, podejście gdzie wyjście  $E_z$  jest zależne od wyjścia  $E_y$ , a także podejście gdzie  $E_z$  oraz  $E_y$  są połączone w jednej Enkoder o współdzielonych warstwach i dwóch wyjściach.

W przypadku cGAN możemy wyróżnić dwa najważniejsze czynniki, które trzeba mieć na uwadze. Pierwszym jest źródło wektora  $y$  podawanego na Generator. W przypadku Dyskryminatora  $y$  jest pobierany ze zbioru treningowego, jednakże w przypadku podawania tego samego wektora na Generator wystąpiła możliwość, że może dojść do niepożdanego przeuczenia modelu. Autorzy artykułu dokonali analizy tego rozwiązania, a także zbadali wydajność metody Bezpośredniej Interpolacji oraz Jądrowego estymatora gęstości. Wynikiem tych badań było stwierdzenie, że dla danej problematyki najlepiej sprawdza się podawanie wektora  $y$  ze zbioru uczącego, możliwość przeuczenia modelu została skomentowana następująco:

*'Jest to możliwe tylko, gdy informacje warunkowe są do pewnego stopnia unikatowa dla każdego obrazu. W tym przypadku, gdzie atrybuty obrazów są binarne, jeden wektor  $y$  może opisać wystarczająco duży i zróżnicowany podzbiór obrazów, zapobiegając nadmiernemu dopasowaniu się modelu do danego  $y$ .'*

Drugim czynnikiem jest warstwa Generatora i Dyskryminatora cGAN na którą podany jest wektor  $y$ . Guim Perarnau oraz jego współpracownicy ustalili, że najlepsze rezultaty otrzymuje się po podaniu wektora  $y$  na warstwę wejściową Generatora oraz pierwszą warstwę konwolucyjną Dyskryminatora.

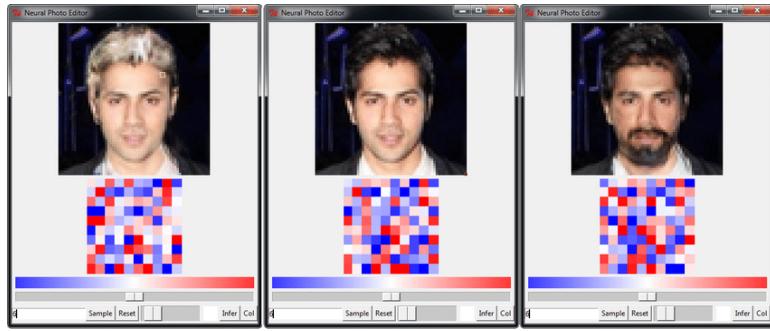
Ważnym spostrzeżeniem z analizy rozwiązania IcGAN jest obecność olbrzymiej ilości różnorodnych rozwiązań opartych na sieciach neuronowych, coraz to nowe architektury zostają wynalezione, aby udoskonalić zastosowania sieci neuronowych do przetwarzania i modyfikowania obrazów.

### 3.4 Neural photo editing

W 2017 roku Andrew Brock, Theodore Lim, J.M. Ritchie and Nick Weston zaprezentowali Neural Photo Editor [2], narzędzie do edytowania obrazu wyposażone w mechanizmy wykrywania kontekstu zmiany. Twórcy opisują swój twór następująco:

*'Interfejs wykorzystujący moc generatywnych sieci neuronowych do wprowadzania dużych, semantycznie spójnych zmian w istniejących obrazach.'*

Użytkowanie wygląda następująco: użytkownik pędzlem o określonym kolorze i rozmiarze maluje na wybranym obrazie, jednak zamiast zmieniać wartości pojedynczych pikseli, interfejs odczytuje kontekst edycji i wprowadza zmiany semantyczne w kontekście żądanej zmiany koloru. Efekt działania interfejsu został przedstawiony na Rysunku 3.4.



Rysunek 3.4: Efekt działania Neural Photo Editor.

Skuteczność NPE (Neural Photo Editor) polega na zastosowaniu IAN (ang. Introspective Adversarial Network), czyli sieci złożonej z połączonych VAE (ang. Variational Autoencoder) [9] oraz GAN, w taki sposób, że dekodująca sieć autoenkodera jest używana jako sieć generująca w GAN. Poprzez przechwytywanie przez model dalekosieżnych zależności, wykorzystanie bloku obliczeniowego bazującego na rozszerzonych splotach o współdzielonych wagach oraz dzięki zastosowaniu ulepszonej generalizacji, udało się osiągnąć dokładną rekonstrukcję obrazu bez strat na jakości detali.

Powstanie NPE utwierdza w przekonaniu, że aktualne możliwości sieci neuronowych do edycji obrazu znacznie przewyższają zwykłe algorytmy pod względem możliwości oraz uzależnienia od wkładu ludzkiego.

## **4 ZIMPLEMENTOWANE ROZWIĄZANIA**

W celu zbadania skuteczności sieci neuronowych jako narzędzi do edycji obrazu należało wybrać przykładowe zagadnienia z tej dziedziny, rozwiązać je z użyciem technik sztucznej inteligencji oraz ocenić ich skuteczność.

W tym rozdziale zostały przedstawione koncepcje rozwiązań wybranych zagadnień oraz ich implementacje.

## **4.1 Automatyczne kolorowanie czarno-białych obrazów**

Problem kolorowania czarno-białych obrazów cieszy się dużym zainteresowaniem z wielu powodów. Od potrzeb kulturowych takich jak możliwość lepszego zwizualizowania oraz zrozumienia przeszłości poprzez kolorowania zdjęć z czasów, kiedy występowały one jedynie w kolorach czerni i bieli, po potrzeby technologiczne takie jak rekonstrukcja filmów oraz poprawa obrazu cyfrowego.

Pomimo braku informacji o kolorze w czarno-białych zdjęciach, ludzie są w stanie określić potencjalne, rzeczywiste barwy obiektów na zdjęciach bazując na treści tych zdjęć oraz swoim doświadczeniu. Można z tego wywnioskować, że zdjęcia te zawierają informacje wystarczające do oszacowania potencjalnych kolorów. Pozwala to założyć, że do tego zagadnienia można skutecznie wykorzystać konwolucyjne sieci neuronowe, które cechują się niezwykłą umiejętnością do rozpoznawania wzorców oraz posiadają wyjątkowe zdolności do adaptacji. Z tego właśnie powodu sieci splotowe zostaną użyte w przedstawionym rozwiązaniu.

### **4.1.1 Podejście**

Rozważając możliwe sposoby pokolorowania czarno-białego zdjęcia można spostrzec, że kiedy niektóre powierzchnie na zdjęciu mają zazwyczaj oczywiste barwy, niebo jest zazwyczaj niebieskie, a trawa zielona, to są też powierzchnie, które posiadają szeroki wachlarz możliwych kolorów, na przykład samochodów może być zarówno czerwony jak i niebieski albo zielony. Z tego powodu celem zaprezentowanego rozwiązania jest niekoniecznie odtworzenie rzeczywistych barw obrazu, a raczej wygenerowanie barw, które mogłyby być barwami rzeczywistymi.

Aby zwiększyć efektywność uczenia wykorzystano przestrzeń barw CIELab. W tej przestrzeni barwę obrazu opisują 3 składowe:

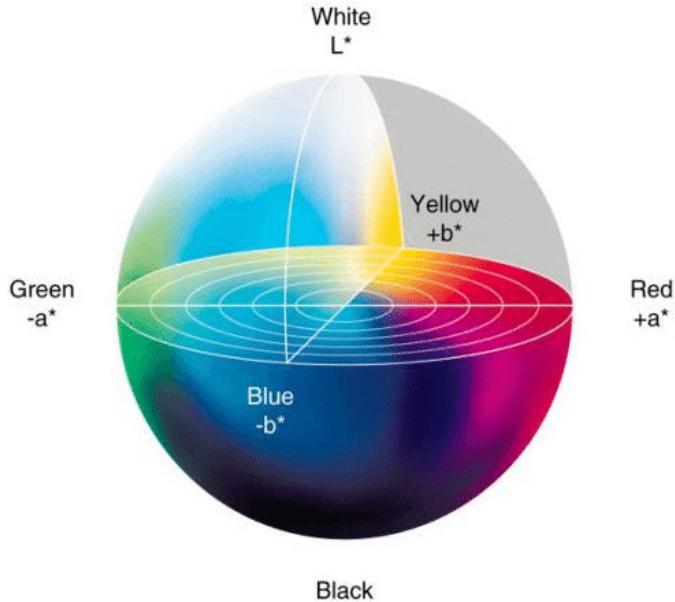
- L -jasność (luminacja)
- A -barwa od zielonej do magenty
- B - barwa od niebieskiej do żółtej

Przestrzeń barw CIELab została przedstawiona na Rysunku 4.1

Zaletą zastosowania CIELab jest fakt, że jest ona najbardziej równomierną przestrzenią barw, co oznacza, że jeśli barwy znajdują się w jednakowej odległości od siebie w tej przestrzeni, to będą one postrzegane jako jednakowo różniące się od siebie. Powinno to zwiększyć skuteczność uczenia sieci oraz zapewnić bardziej realistyczne kolorowanie.

Składowa  $L$ , jako, że jest identyczna dla obrazu kolorowego jak i czarno-białego, stanowi w tym przypadku wejście sieci, na jej podstawie sieć odtwarza składowe  $A$  oraz  $B$ , które reprezentują przewidziane kolory dla obrazu wejściowego.

Jako rozwiązanie podanej problematyki wpierw został oceniony autorski model podstawowy, na jego podstawie zostało przeprowadzone porównanie skuteczności różnych konfiguracji, w których uczony był model. Do elementów poddanych testom należą algorytm optymalizacyjny, funkcja straty, funkcja aktywacją oraz sposób przetwarzania wstępnie danych treningowych.



Rysunek 4.1: Przestrzeń barw CIELab.

#### 4.1.2 Model podstawowy

Opracowany przez nas model jest to FCN. Konwolucyjna część sieci składa się z 12 warstw splotowych mających na celu nauczyć się mapować składową wejściową  $L$  na wyjściowe składowe  $A$  i  $B$ . Składowe wyjściowe muszą mieć takie same wymiary jak składowa wejściowa, co oznacza, że kluczowym było odpowiednie dobranie parametrów warstw takich jak *padding* (pol. otoczka), *stride* (pol. krok) oraz wielkość filtrów. Pełna architektura sieci została przedstawiona w Tabeli 1. Pierwsza warstwa konwolucyjna rozkłada wejściowy kanał na 32 kanały, co pozwala wyciągnąć z niego jak najwięcej informacji o cechach obrazu. Warstwa ta ma wielkość filtra  $3 \times 3$ , tak więc aby zachować wymiar kanałów zostały zastosowane parametry  $padding = 1$  oraz  $stride = 1$ .

Kolejne trzy warstwy ekstraktyują z wejściowych 32 kanałów najbardziej istotne cechy związane z powiązaniem treści obrazu z szacowanym kolorem jego powierzchni. Warstwy te na swoje wyjście przekazują po 32 kanały zawierające wykryte powiązanie pomiędzy pikselami kanałów wejściowych.

Warstwa piąta rozciąga wejściowe 32 kanały na 64 kanały, dzięki temu kolejne 2 warstwy przyjmujące na wejście te 64 kanały i przekazujące je na wyjście są w stanie wydobyć z obrazu cechy o większym poziomie abstrakcji, co znacznie zwiększa skuteczność działania sieci.

Warstwa ósma ogranicza ilość kanałów w sieci z 64 do 32 wyciągając z nich cechy najbardziej przydatne do rozwiązyania danej problematyki. Kanały te są następnie ponownie przetwarzane przez warwę z wielkością filtra  $3 \times 3$ , co ma służyć agregacji rozłożonych cech w bardziej spójną całość, która może być już składana w pożądane wyjście.

Nr	Warstwa	Rozmiar filtra	Stride	Padding	Batch Normalization	Fun. aktywacji	Ilość kanałów wej./wyj.
1	Splotowa	3x3	1	1	Tak	RELU	1/32
2	Splotowa	3x3	1	1	Tak	RELU	32/32
3	Splotowa	3x3	1	1	Tak	RELU	32/32
4	Splotowa	3x3	1	1	Tak	RELU	32/32
5	Splotowa	3x3	1	1	Tak	RELU	32/64
6	Splotowa	3x3	1	1	Tak	RELU	64/64
7	Splotowa	3x3	1	1	Tak	RELU	64/64
8	Splotowa	3x3	1	1	Tak	RELU	64/32
9	Splotowa	3x3	1	1	Tak	RELU	32/32
10	Splotowa	1x1	1	0	Tak	RELU	32/32
11	Splotowa	1x1	1	0	Tak	RELU	32/32
12	Splotowa	1x1	1	0	Nie	-	32/2

Tabela 1: Architektura modelu podstawowego.

Kolejne dwie warstwy w sieci są to warstwy konwolucyjne o wielkości filtru 1x1, odpowiadające one warstwom gęstym i mają na celu przekonwertowanie wartości funkcji aktywacji z poprzednich warstw na wartości kolorów odpowiadających pikseli w przestrzeni barw CIELab. Ostatnia warstwa, również z filtrem o wielkości 1x1 zwija 32 kanały otrzymywane na wejściu do 2 kanałów odpowiadających składowym  $A$  oraz  $B$ , które stanowią pożądany rezultat działania sieci.

Po wszystkich, oprócz ostatniej, warstwach konwolucyjnych znajdują się dodatkowo warstwa BatchNorm oraz warstwa funkcji aktywacji RELU mające na celu ustabilizować proces uczenia oraz zwiększyć jego efektywność.

#### 4.1.3 BatchNorm

Warstwa Batch Normalization została przedstawiona w 2015 roku przez S. Ioffe oraz C. Szegedy jako odpowiedź na problem zmieniającej się podczas uczenia dystrybucji wartości wejść każdej z warstw sieci [10]. Ma ona na celu usprawnić i ustabilizować trening sieci poprzez normalizację wartości podawanych na funkcje aktywacji. Zmienna dystrybucja tych wartości znacznie spowalnia i utrudnia proces uczenia poprzez potrzebę przemyślanego inicjowania wag sieci w celu zwiększenia prawdopodobieństwa nakierowania modelu na pożądane rozwiązania w trakcie procesu uczenia oraz przez konieczność używania mniejszych wartości współczynnika uczenia, aby przeciwdziałać problemom zanikającego oraz wybuchającego gradientu [11].

Problemy te zostały jużauważone i opisane w 1994 roku przez Y. Bengio oraz jego współpracowników [11]. Dowodzą oni, że:

*'Metoda gradientu prostego staje się coraz bardziej nieefektywna, gdy rośnie czasowy zakres zależności'*

Wskazują także, że problemy powstają podczas treningu DNN w fazie wstecznej propagacji błędu, kiedy to gradient pochodzący z głębszych warstw przechodzi wielokrotnie przez operacje mnożenia macierzowego. Jeśli wartość gradientu jest niewielka, to z każdą operacją mnożenia staje się jeszcze mniejsza, aż maleje do takich wartości, które nie umożliwiają modelowi uczenia się, a jeśli wartość ta jest wysoka to, wraz z przechodzeniem przez kolejne warstwy, rośnie jeszcze bardziej co przy bardzo dużych wartościach może doprowadzić do destabilizacji procesu uczenia. Są to zjawiska zdecydowanie niepożądane i z tego powodu powstało wiele rozwiązań, aby im przeciwdziałać takich jak ograniczanie maksymalnej wartości gradientu (ang. gradient clipping) albo zastosowanie warstw BatchNorm.

Zastosowanie warstw BatchNorm sprawia, że podczas uczenia metodą mini-batch (pol. małych paczek) każda paczka jest normalizowana w sposób zapewniający zerową wartość średnią oraz równą jedności wariancję na przestrzeni wszystkich kanałów wejściowych. Zaletą takiego podejścia jest poprawienie przepływu korygującego gradientu przez kolejne warstwy sieci podczas fazy wstecznej propagacji błędu. Ponadto warstwy BatchNorm zapewniają większą odporność sieci na niekorzystnie zainicjowane wagи początkowe modelu.

Użycie tych warstw w modelu podstawowym tuż za warstwami RELU pozwoliło uzyskać bardziej korzystną zbieżność modelu oraz lepsze rezultaty końcowe. Ocenione zostało też rozwiązania, w którym warstwy BatchNorm znajdują się przed warstwami funkcji aktywacji, lecz dało ono gorsze rezultaty, niż podejście wspomniane jako pierwsze.

#### 4.1.4 Dropout

W trakcie pracy nad ostatecznym modelem podstawowym sprawdzona została skuteczność zastosowania warstw Dropout [12]. Warstwy te w trakcie treningu dezaktywują część neuronów aby przeciwdziałać efektowi przeuczania się sieci oraz zapewniać wydajny sposób łączenia wielu różnych architektur sieci stworzonych do jednego celu w jednolitą całość o skuteczności większej niż poszczególne sieci osobno.

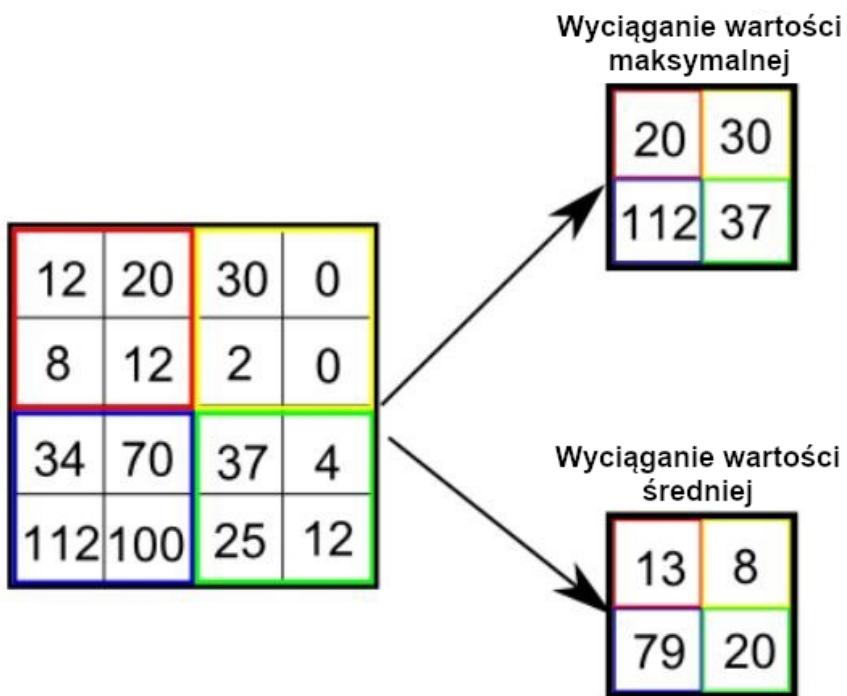
Wybór neuronów, dla których w danej iteracji treningu nie zostaną zaktualizowane wagи odbywa się z pewnym prawdopodobieństwem określonym podczas inicjalizacji modelu. Dezaktywowanie neuronów można interpretować jak przerywanie tymczasowo wszystkich połączeń danego neuronu, zarówno wejściowych jak i wyjściowych. Takie działania są równoznaczne z wyselekcjonowaniem z modelu mniejszej sieci i trenowaniu wyłącznie jej w aktualnej iteracji. Wagi tej podsieci są wtedy współdzielone z modelem źródłowym. Jako rezultat uzyskuje się model o znacznie ulepszonych zdolnościach generalizacji.

Zastosowanie tych warstw w modelu nie przyniosło wyraźnego polepszenia rezultatów sieci. W przypadku danej problematyki oraz obranego podejścia do jej rozwiązania, przeuczenie sieci nie stanowi wyraźnego zagrożenia, a zastosowanie Dropout wiąże się z utratą części informacji kluczowych do odpowiedniego generowania kolorów dla wejściowych obrazów. Model powinien nauczyć się jak największej różnorodności kolorów, a Dropout przeciwdziała uczeniu się nadmiernej ilości cech przez sieć, co wpływa niekorzystnie na otrzymywane rezultaty końcowe. Podczas testów skuteczności tych warstw zostały one umieszczone za warstwami BatchNorm. Wizualizacja skutków tej decyzji wraz z rozważeniem pozostałych czynników wpływających na efektywność sieci znajdują się w punkcie 4.1.14 związanym z rezultatami modelu podstawowego.

#### 4.1.5 Modyfikacja rozdzielczości

W modelach FCN powszechnie stosuje się różne metody zmiany rozdzielczości kanałów przeprowadzających przez sieć, najczęściej są to operacje poolingu mające na celu zredukować przestrzenną wielkość reprezentacji cech wyciągniętych z obrazu poprzez wyciągnięcie najbardziej istotnych wartości funkcji aktywacji z określonych obszarów reprezentacji w celu zredukowania rozmiaru sieci, a co za tym idzie, zmniejszenia ilości obliczeń koniecznych do wykonania przez sieć.

Ponadto pooling wspomaga adaptację modelu do zmiennego położenia kluczowych wzorów rozpoznawanych przez sieć na obrazie wejściowym. Cechą ta zwana jest niezmiennością od translacji (ang. transaltion invariance). Jest ona rezultatem dokonywania operacji, takich jak wyliczanie wartości maksymalnej z poszczególnych obszarów, dzięki którym zmienne położenie wartości selekcjonowanej, a co tym idzie, zmienne położenie ekstraktowanej cechy w obrębie danego obszaru nie wpływa na końcową postać reprezentacji przestrzennej obrazu wejściowego po przejściu przez warstwę poolingu. Przykładowe operacje poolingu zostały przedstawione na Rysunku 4.2.



Rysunek 4.2: Przykładowe operacje warstwy poolingu

W modelu końcowym warstwy poolingu nie zostały zastosowane, aby uniknąć utraty kluczowych informacji przestrzennych koniecznych do właściwego wygenerowanie możliwych barw obrazu.

#### 4.1.6 Wykorzystywany zbiór treningowy

Do uczenia modelu został wykorzystany zbiór danych CIFAR-10 stworzony przez A. Krizhevsky oraz zaprezentowany w 2009 roku [13]. Składa się on z 60000 obrazów w przestrzeni kolorów RGB o rozdzielczości 32 x 32 piksele. Spośród tych obrazów 10000 z nich zostało wykorzystanych jako zbiór walidacyjny do śledzenia skuteczności treningu modelu. Przykładowe obrazy ze zbioru danych zostały przedstawione na Rysunku 4.3.



Rysunek 4.3: Przykładowe obrazy z CIFAR-10

Zaletą CIFAR-10 jest niewielka rozdzielcość obrazów co pozwala na mniej złożony model oraz szybszy proces uczenia, który można skutecznie przeprowadzić nawet przy ograniczonych możliwościach obliczeniowych. Ponadto zbiór ten składa się z obrazów dzielących się na 10 klas, tak mała różnorodność klas, a co za tym idzie, stosunkowo niewielka ilość możliwych obiektów pojawiających się na zdjęciach powinna ułatwić sieci wyuczenie się właściwych barw dla identyfikowanych powierzchni. Z drugiej strony mała ilość klas może niekorzystnie wpływać na umiejętności generalizacji modelu, jednakże obrazy z CIFAR-10 przedstawiają zróżnicowane otoczenia zawierające obiekty nie kwalifikujące się do żadnej klasy, co powinno umożliwić wyuczenia się generalizacji przez sieć.

#### 4.1.7 Przetwarzanie wstępne danych

Przed podaniem na wejście sieci obrazy uczące były wpierw poddawane przetwarzaniu wstępnemu mającemu na celu doprowadzić do szybszego oraz stabilniejszego treningu modelu. Przetwarzanie wstępne jest kluczowe, gdyż wartości o jakie aktualizowane są wagie neuronu zależą w dużej mierze od wartości wejść tego neuronu. W przypadku gdy przedziały wartości tych wejść nie są jednolite, to może wystąpić duża różnica w tempie aktualizacji wag sieci, niektóre wagie będą zmieniane o wiele szybciej niż inne co może spowodować destabilizację treningu. Przeskalowanie wszystkich wartości wejściowych do jednakowych przedziałów, o niewielkiej wartości maksymalnej i minimalnej oraz wartości średniej zbliżonej do zera zmniejsza możliwość wystąpienia tego problemu oraz sprzyja ujednoliceniu tempa uczenia się przez sieć rozpoznawania różnych cech. Ponadto brak przeskalowania wejść może doprowadzić do zjawisk wybuchającego oraz zanikającego gradientu.

W ramach badania rozwiązania przetestowane zostały różne metody przetwarzania wstępnego zarówno danych wejściowych jak i pożądanej odpowiedzi:

- Normalizacja danych na przestrzeni całego zbioru danych z użyciem wartości maksymalnej oraz minimalnej całego zbioru, tak aby wartości pikseli danej składowej dla każdego obrazu zawierały się w przedziale od -0.5 do 0.5.
- Standaryzacja danych na przestrzeni całego zbioru danych z użyciem wartości średniej oraz odchylenia standardowego całego zbioru, tak aby uzyskać wartość średnią równą w przybliżeniu zero oraz jednostkowe odchylenie standardowe.
- Zastosowanie rozmycia gaussowskiego(ang. Gaussian blur) o różnej wielkości filtru Gaussowskiego.
- Przepuszczenie składowej przez filtr Gaussa o różnej wielkości filtru.

Rozmycie gaussowskie, zwane także wygładzaniem gaussowskim (ang. Gaussian smoothing), jest to operacja polegająca na modyfikacji obrazu z użyciem filtru Gaussa. Stosuje się je w celu rozmycia detali na przetwarzanym obrazie, a także by ograniczyć ilość występujących na nim zakłóceń oraz szumów. Jest ono powszechnie stosowane w fazie przetwarzania wstępnie danych graficznych. W praktyce operacja ta sprowadza się do dokonywania splotu kolejnych fragmentów obrazu z funkcją Gaussa. Zastosowanie jej na danych wejściowych miało na celu zmniejszenie znaczenia detali na obrazie oraz ułatwienie sieci nauczenia się rozróżniania rozmaitych powierzchni oraz wzorców.

Wymienione metody przetwarzania były stosowane w różnych połączeniach oraz konfiguracjach zarówno na składowej  $L$ , jak i składowych  $A$  oraz  $B$ , a uzyskane wyniki opisane zostały w punkcie 4.1.14

#### *4.1.8 Przetwarzanie końcowe danych*

Jeśli w trakcie procesu uczenia były stosowane zabiegi przetwarzania wstępnego danych wejściowych to podczas testowania modelu dane testowe również muszą być przetworzone w ten sam sposób, aby zapewnić poprawną pracę sieci. To samo dotyczy się danych wyjściowych sieci, jeśli podczas treningu z nadzorem pożądane wyjście było w pewien sposób przetworzone wstępnie to sieć uczy się odtwarzać te wyjście tak samo przetworzone, aby uzyskać oczekiwany efekt końcowy należy dokonać operacji odwrotnych do tych zastosowanych w fazie przetwarzania wstępnego. Przykładowo jeśli pożądana odpowiedź w procesie uczenia była normalizowana to podczas testów sieci jej wyście należy zdenormalizować, aby uzyskać oczekiwany rezultat. Jednakże my dla rozważanej problematyki proponujemy metodę alternatywną.

Polega ona na uwydatnianiu kolorów wygenerowanych przez sieć dla wysokich wartości natświetlenia - składowej  $L$ . Może być ona stosowana niezależnie od metody przetwarzania wstępnego danych wejściowych. Algorytm polega na przeskalowaniu pikseli składowych  $A$  i  $B$  tak, aby ich wartości mogły pokryć cały dostępny przedział wartości, ale dla danego piksela jego przedział jest ograniczony proporcjonalnie do stosunku wartości tego piksela w składowej  $L$  do maksymalnej możliwej wartości pikseli składowej  $L$ .

Dla przykładu założmy, że wartość danego pikselu dla składowej  $A$  wynosi 70, dla składowej  $B$  wynosi -20, a dla składowej  $L$  80. Przedział wartości składowych  $A$  i  $B$  jest od -127 do 128, a składowej  $L$  od 0 do 100. W kolejnym kroku znajdowana jest maksymalna absolutna wartość składowych  $A$  i  $B$  dla aktualnego obrazu. Założymy, że dla  $A$  wynosi ona 90, a dla  $B$  60. Następnie piksele składowych  $A$  i  $B$  są dzielone przez swoje maksymalne wartości, a następnie rozciagnięte na cały dostępny przedział przez pomnożenie przez odpowiedni czynnik, czynnik ten jest z przedziału od 0 do 127 i jest wprost proporcjonalny do stosunku aktualnego piksela składowej  $L$  do maksymalnej wartości  $L$ , oznacza to, że jeśli dany piksel  $L$  jest równy 100 to czynnik jest równy 127, a jeśli dany piksel  $L$  jest równy 50 to wartość czynnika znajduje się w połowie swojego przedziału i równy jest 63,5.

Dla podanych założeń otrzymujemy następujące nowe wartości piksela składowej  $A$  ( $p_n^A$ ) i piksela składowej  $B$  ( $p_n^B$ ):

$$p_n^A = \frac{70}{90} * 127 * \frac{80}{100} = 79.02 \quad (1)$$

$$p_n^B = \frac{-20}{60} * 127 * \frac{80}{100} = -33.87 \quad (2)$$

Algorytm konwertowanie pikseli danej składowej można przedstawić wzorem:

$$S_{i,j}^n = \frac{S_{i,j}^p}{\max\{\text{abs}\{S\}\}} * 127 * \frac{L_{i,j}}{100} \quad (3)$$

Gdzie:

- $S$  - Konwertowana składowa będąca dwuwymiarowa macierzą pikseli.
- $S_{i,j}^n$  - Nowa wartość piksela  $(i, j)$  dla danej składowej.
- $S_{i,j}^p$  - Stara wartość piksela  $(i, j)$  dla danej składowej.
- $L_{i,j}$  - Wartość piksela  $(i, j)$  składowej jasności.

Zastosowanie powyższego algorytmu pozwoliło uzyskać bardziej zadawalające rezultaty działania modelu poprzez faworyzowanie kolorów tworzonych przez sieć dla powierzchni o dużej wartości jasności, jako, że wysoka jasność oferuje bardziej intensywne barwy. Kolorystyka powierzchni ciemnych pozostaje przytłumiona, jako, że brak naświetlenia ogranicza natężenie kolorów.

#### 4.1.9 Augmentacja danych

W celu skutecznego treningu modelu potrzebny jest odpowiednio duży i różnorodny zbiór treningowy. W obliczu problemu niewystarczającej ilości danych stosuje się metody zwane augmentacją danych pozwalające poszerzyć zbiór obrazów uczących poprzez dodawanie nowych informacji do obrazów będących podstawą zbioru tworząc w ten sposób nowe obrazy, które mogą być wykorzystane w treningu. Augmentacja danych przeciwdziała uczeniu się przez sieć nieistotnych wzorów i cech takich jak orientacja obiektu, jego umiejscowienie albo rozmiar. Dzięki temu model jest w stanie dogłębniej analizować obrazy wejściowe ucząc się rozpoznawać cechy o coraz to większym poziomie abstrakcji co przekłada się na o wiele lepiej rozwiniętą zdolność modelu do generalizacji danej problematyki.

Do augmentacji stosuje się proste przekształcenia obrazu takie jak:

- Rotacja obrazu o wybrany kąt.
- Odbicie obrazu względem osi pionowej.
- Obcinanie skrajnych fragmentów obrazów (ang. crop).
- Zmiana odcienia oraz saturacji barw obrazu.
- Przybliżanie albo oddalanie treści obrazu.
- Modyfikacja rozdzielczości obrazu poprzez jego rozciąganie lub ściskanie.
- Tworzenie niewielkich ubytków w obrazach w losowych miejscach (ang. coarse dropout).

Należy jednak pamiętać, że nie wszystkie metody augmentacji nadają się do każdego zbioru danych, kluczowe jest, aby wybrać takie operacje, które poszerzą zbiór uczący o dane niosące znaczące oraz sensowne informacje patrząc z punktu wybranej problematyki oraz nie przesłaniają wzorców kluczowych do wyuczenia przez sieć. W przypadku zagadnienia kolorowania czarno-białych obrazów kluczową informacją jest kolor analizowanej powierzchni oraz jej cechy charakterystyczne, z tego powodu, do treningu modelu podstawowego nie zostały wykorzystane żadne metody augmentacji wpływające na barwy albo jasność obrazów. Wykluczone zostały również takie metody jak tworzeniu ubytków w obrazach, gdyż powoduje to niepotrzebną utratę informacji. W związku z tym, że wykorzystany zbiór CIFAR-10 posiada dużą ilość obrazów to w procesie uczenia została wykorzystana jedynie augmentacja poprzez odbicie obrazu względem osi pionowej z prawdopodobieństwem równym 50%.

#### 4.1.10 Funkcje kosztów

Kluczem do właściwego funkcjonowania modelu jest wybór odpowiedniej funkcji kosztu. W przypadku problematyki kolorowania czarno-białych obrazów ważne jest, aby wybrana funkcja kosztu uwzględniała specyficzną naturę problemu, gdzie dla niektórych przypadków właściwych jest wiele odpowiedzi. Jako przykład można podać taki obiekt jak samochód, który może być zarówno zielony, czerwony jak i żółty, każdy z tych kolorów powinien być oceniony jako właściwy, a wartość błędu wyliczona z użyciem funkcji kosztu dla tak wybranych przez sieć kolorów powinna odpowiednio to wskazywać. W ramach poszukiwań najbardziej odpowiedniej funkcji kosztu przetestowane zostały następujące funkcje.

1. MSELoss (ang. Mean Squared Error Loss) - koszt oparty na błędzie średniokwadratowym przedstawiony funkcją:

$$Koszt = \frac{1}{n} \sum_{i=0}^n (x_i - y_i)^2 \quad (4)$$

Po dopasowaniu równania MSELoss do rozważanej problematyki otrzyma się:

$$Koszt = \frac{1}{n} \frac{1}{m} \sum_{i=0}^n \sum_{j=0}^m ((A_{i,j}^P - A_{i,j}^R)^2 + (B_{i,j}^P - B_{i,j}^R)^2) \quad (5)$$

2. L1Loss zwany także MAELoss (ang. Mean Absolute Error Loss) - koszt oparty na średnim błędzie bezwzględnym dany funkcją:

$$Koszt = \frac{1}{n} \sum_{i=0}^n |x_i - y_i| \quad (6)$$

Równanie L1Loss dla rozważanego zagadnienia:

$$Koszt = \frac{1}{n} \frac{1}{m} \sum_{i=0}^n \sum_{j=0}^m (|A_{i,j}^P - A_{i,j}^R| + |B_{i,j}^P - B_{i,j}^R|) \quad (7)$$

3. SmoothL1Loss - zwany także *Huber loss*, odmiana L1Loss przedstawiona w 2015 roku przez R. Girshick [14]. Jej zaletami są mniejsza czułość na elementy odstające (ang. outlier) i zmniejszona szansa na wystąpienie zjawiska eksplodującego gradientu. SmoothL1Loss przedstawiony jest funkcją:

$$Koszt = \frac{1}{n} \sum_{i=0}^n z_i \quad (8)$$

gdzie  $z_i$  dane jest:

$$z_i = \begin{cases} 0.5(x_i - y_i)^2, \text{ jeśli } |x_i - y_i| < 1 \\ |x_i - y_i| - 0.5, \text{ w pozostałych przypadkach} \end{cases} \quad (9)$$

Zastosowanie SmoothL1Loss do modelu podstawowego da następującą funkcję:

$$Koszt = \frac{1}{n} \sum_{i=0}^n (z_i + k_i) \quad (10)$$

gdzie  $z_i$  dane jest:

$$z_i = \begin{cases} 0.5(A_{i,j}^P - A_{i,j}^R)^2, \text{ jeśli } |A_{i,j}^P - A_{i,j}^R| < 1 \\ |A_{i,j}^P - A_{i,j}^R| - 0.5, \text{ w pozostałych przypadkach} \end{cases} \quad (11)$$

a  $k_i$  dane jest:

$$z_i = \begin{cases} 0.5(B_{i,j}^P - B_{i,j}^R)^2, \text{ jeśli } |B_{i,j}^P - B_{i,j}^R| < 1 \\ |B_{i,j}^P - B_{i,j}^R| - 0.5, \text{ w pozostałych przypadkach} \end{cases} \quad (12)$$

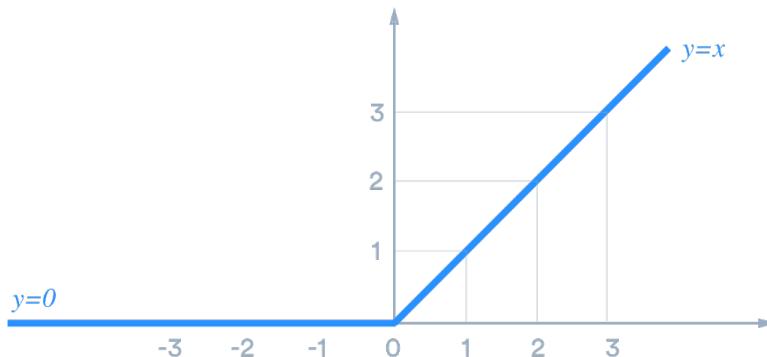
Gdzie:

- $x$  są to składowe  $A$  oraz  $B$  przewidziane przez sieć.
- $y$  to rzeczywiste składowe  $A$  i  $B$ .
- $A_{i,j}^R$  i  $B_{i,j}^R$  są to rzeczywiste wartości składowych  $A$  i  $B$  dla pikselu obrazu o współrzędnych  $(i, j)$ .
- $A_{i,j}^P$  i  $B_{i,j}^P$  są to wartości składowych  $A$  i  $B$  przewidziane przez sieć dla pikselu obrazu o współrzędnych  $(i, j)$ .

Szczegółowy opis skutków zastosowanie poszczególnych funkcji kosztów umieszczony został w punkcie 4.1.14.

#### 4.1.11 Funkcje aktywacji

W wyborze odpowiedniej funkcji aktywacji kierowaliśmy się badaniami przeprowadzonymi przez B. Xu oraz jego współpracowników [15]. Testowali oni skuteczność takich funkcji jak ReLU, Leaky ReLU (pol. przepuszczające ReLU), PReLU (ang. Parametric ReLU) oraz RReLU (ang. Randomized Leaky ReLU) w zagadnienu klasyfikacji treści obrazów. Bazując na otrzymanych przez nich rezultatach zdecydowaliśmy się wykorzystać w rozważanej problematyce funkcję aktywacji ReLU, przedstawioną na Rysunku 4.4



Rysunek 4.4: Funkcja aktywacji ReLU

Funkcja ta posiada wiele zalet takich jak przyspieszenie procesu zbiegania się stanu sieci do stanu pożdanego poprzez brak ograniczeń na maksymalną wartość funkcji oraz niska złożoność obliczeniowa. Ponadto funkcja ta, w związku z zerową wartością dla ujemnych argumentów, zapewnia aktywację neuronów modelu tylko wtedy, gdy analizują one wzorce kluczowe dla nich samych oraz danego zagadnienia, zwiększa to odporność modelu na szумy wejściowe oraz przeuczanie, a także poprawia zdolności predykcyjne sieci.

Jednakże stosowanie ReLU tworzy zagrożenie blokowanie się procesu uczenia, jeśli dojdzie do sytuacji, w której wagi modelu dojdą do stanu, gdzie wartość aktywacji będzie zbliżona do zera. Spowoduje to zerową wartość gradientu podczas fazy wstępnej propagacji błędu powodując wstrzymanie się procesu aktualizowania wag modelu, a w rezultacie, nieefektywny proces uczenia. Pomimo to jednak zostaliśmy przy wyborze funkcji ReLU wierząc, że pozostałe zabiegi takie jak zastosowanie warstw BatchNorm pozwolą zminimalizować niepożądane efekty tego zjawiska.

#### 4.1.12 Algorytmy optymalizacyjne

Podczas planowania treningu modelu należy zdecydować się na odpowiedni algorytm optymalizacyjny, właściwa decyzja pozwala uniknąć zatrzymania się procesu uczenia w lokalnych minimach co zwiększa końcową dokładność oraz skuteczność modelu. Podczas badań przetestowane zostały różne algorytmy optymalizacyjne, a uzyskane rezultaty zostały szczegółowo opisane oraz porównane w punkcie 4.1.14.

Wykorzystane zostały następujące algorytmy:

- Adam (ang. Adaptive Moment Estimation)
- Adagrad (ang. Adaptive Gradient Algorithm)

- SGD (ang. Stochastic Gradient Descent)

Przeprowadzone zostały także poszukiwania najbardziej odpowiednich hiperparametrów dla wymienionych algorytmów w celu osiągnięcia jak największej ich skuteczności w rozwiązaniu rozważanej problematyki.

#### *4.1.13 Trening*

Model trenowaliśmy paczkami o wielkości 128 obrazów, cała epoka składała się z 50000 obrazów. Przed każdą epoką zbiór treningowy był przetasowywany, aby przeciwdziałać przeuczaniu się sieci. Aby znaleźć najbardziej zadowalające rozwiązanie przetestowane zostały różne konfiguracje treningowe, możliwe zmienne elementy konfiguracji to algorytmy optymalizacyjne opisane w punkcie 4.1.12, funkcje kosztów opisane w punkcie 4.1.10, rodzaje przetwarzania wstępne opisane w punkcie 4.1.7 oraz skutki zastosowanie takich warstw jak Dropout (punkt 4.1.4) oraz BatchNorm (punkt 4.1.3). Jako funkcja aktywacji wybrana została funkcja ReLU. Dla różnych konfiguracji została także przedstawiona charakterystyka porównawcza.

#### *4.1.14 Rezultaty*

Najbardziej satysfakcjonujące rezultaty uzyskane przez model zostały przedstawione na Rysunku 4.5



(a) Obraz zewnętrzny



(b) Obraz zewnętrzny



(c) Obraz z datasetu

Rysunek 4.5: Rezultaty.

## **5 PODSUMOWANIE**

## BIBLIOGRAFIA

- [1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio: *Generative Adversarial Networks*, arXiv ('2014)
- [2] Andrew Brock, Theodore Lim, J.M. Ritchie, Nick Weston: *NEURAL PHOTO EDITING WITH INTROSPECTIVE ADVERSARIAL NETWORKS*, arXiv ('2017)
- [3] Tianyi Liu, Shuang Sang Fang, Yuehui Zhao, Peng Wang, Jun Zhang: *Implementation of Training Convolutional Neural Networks*, arXiv ('2015), s.2.
- [4] Ian Goodfellow , Yoshua Bengio, Aaron Courville: *Deep Learning*, ('2016), s.342.
- [5] Richard Zhang, Phillip Isola, Alexei A. Efros: *Colorful Image Colorization*, arXiv ('2016)
- [6] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge: *Image Style Transfer Using Convolutional Neural Networks*, IEEE ('2016)
- [7] Karen Simonyan, Andrew Zisserman: *Very Deep Convolutional Networks For Large-Scale Image Recognition*, ('2016)
- [8] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, Jose M. Álvarez: *Invertible Conditional GANs for image editing*, arXiv ('2016)
- [9] Diederik P. Kingma, Max Welling: *Auto-Encoding Variational Bayes*, arXiv ('2014)
- [10] Sergey Ioffe, Christian Szegedy: *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, arXiv ('2015)
- [11] Y. Bengio, P. Simard, P. Frasconi: *Learning long-term dependencies with gradient descent is difficult*, IEEE ('1994)
- [12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov: *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, Journal of Machine Learning Research ('2014)
- [13] A. Krizhevsky: *Learning Multiple Layers of Features from Tiny Images*, Master's thesis, Department of Computer Science, University of Toronto ('2009)
- [14] R. Girshick: *Fast R-CNN*, arXiv ('2015)
- [15] B. Xu, N. Wang, T. Chen, M. Li: *Empirical Evaluation of Rectified Activations in Convolution Network*, arXiv ('2015)

## **Spis rysunków**

2.1	Struktura DNN - źródło: <a href="https://towardsdatascience.com/building-a-convolutional-neural-network-male-vs-female-50347e2fa88b">https://towardsdatascience.com/building-a-convolutional-neural-network-male-vs-female-50347e2fa88b</a> . . . . .	8
2.2	Przykładowa struktura CNN - źródło: <a href="https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html">https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html</a> . . . . .	9
2.3	Segmentacja semantyczna - źródło: <a href="https://devblogs.nvidia.com/image-segmentation-using-digits-5/">https://devblogs.nvidia.com/image-segmentation-using-digits-5/</a> . . . . .	10
3.1	Efekt kolorowanie czarno-białych zdjęć przez wytrenowany model - źródło: [5] . . . . .	12
3.2	Obrazy będące kombinacją treści zdjęcia ze stylami kilku znanych dzieł sztuki - źródło: [6] . . . . .	13
3.3	Obrazy generowane przez IcGAN - źródło: [8] . . . . .	14
3.4	Efekt działania Neural Photo Editor - źródło: [2] . . . . .	16
4.1	Przestrzeń barw CIELab - źródło: <a href="https://www.flickr.com/photos/greenmambagreenmamba/4236391637">https://www.flickr.com/photos/greenmambagreenmamba/4236391637</a> . . . . .	19
4.2	Przykładowe operacje warstwy poolingu - źródło: Rysunek własny . . . . .	22
4.3	Przykładowe obrazy z CIFAR-10 - źródło: Rysunek własny . . . . .	23
4.4	Funkcja aktywacji ReLU - źródło: <a href="https://pytorch.org/docs/stable/_images/ReLU.png">https://pytorch.org/docs/stable/_images/ReLU.png</a> . . . . .	28
4.5	Rezultaty - źródło: Rysunek własny na podstawie kolejno: <a href="https://fr.m.wikipedia.org/wiki/Fichier:An_F-A-18C_Hornet_launches_from_the_flight_deck_of_the_conventionallyPowered_aircraft_carrier.jpg">https://fr.m.wikipedia.org/wiki/Fichier:An_F-A-18C_Hornet_launches_from_the_flight_deck_of_the_conventionallyPowered_aircraft_carrier.jpg</a> , <a href="https://pl.wikipedia.org/wiki/Plik:PL_Bagno_Całowanie_2.jpg">https://pl.wikipedia.org/wiki/Plik:PL_Bagno_Całowanie_2.jpg</a> , [13] . . . . .	30

## **Spis tabel**

1	Architektura modelu podstawowego. . . . .	20
---	---	----