



POLITECHNIKA  
LUBELSKA  
WYDZIAŁ PODSTAW  
TECHNIKI

---

**KIERUNEK: MATEMATYKA**



**Praca magisterska**

Przegląd autoenkoderów stosowanych w nienadzorowanym uczeniu  
maszynowym

*An overview of autoencoders used in unsupervised machine learning*

*Praca wykonana pod kierunkiem:*  
dra Dariusza Majerka

*Autor:*  
Alicja Hołowiecka  
nr albumu: 89892

**Lublin 2022**



# Spis treści

<b>Wstęp</b> . . . . .	5
<b>Rozdział 1. Podstawy teoretyczne</b> . . . . .	7
1.1. Sztuczne sieci neuronowe . . . . .	7
1.1.1. Funkcje aktywacji . . . . .	10
1.1.2. Funkcje straty . . . . .	11
1.2. Sieci splotowe . . . . .	12
1.2.1. Czym jest sieć splotowa . . . . .	12
1.2.2. Elementy sieci splotowej . . . . .	13
1.3. Autoenkodery . . . . .	22
1.3.1. Czym jest autoenkoder . . . . .	22
1.3.2. Rodzaje autoenkoderów . . . . .	23
1.3.3. Zastosowania autoenkoderów . . . . .	38
<b>Rozdział 2. Część praktyczna</b> . . . . .	45
2.1. Prosty autoenkoder . . . . .	45
2.2. Autoenkoder splotowy . . . . .	48
2.3. Autoenkoder odszumiający . . . . .	50
2.4. Wyszukiwanie obrazu . . . . .	52
2.5. Wykrywanie anomalii przy użyciu autoenkodera . . . . .	54
2.6. Generowanie obrazów przy użyciu autoenkodera wariancyjnego . . . . .	57
<b>Podsumowanie i wnioski</b> . . . . .	63
<b>Bibliografia</b> . . . . .	65
<b>Spis rysunków</b> . . . . .	67
<b>Spis tabel</b> . . . . .	71
<b>Załączniki</b> . . . . .	73
<b>Streszczenie (Summary)</b> . . . . .	75



## Wstęp

Jeszcze niedawno komputery nie były w stanie wykonywać wielu zadań, które dla człowieka wydają się trywialne, jak na przykład rozpoznanie obiektu znajdującego się na zdjęciu. W ostatnich latach następuje jednak coraz większy rozwój w dziedzinie sztucznych sieci neuronowych, inspirowanych do pewnego stopnia biologicznym mózgiem człowieka. Zainteresowanie sztucznymi sieciami neuronowymi jest potęgowane przez dostępność ogromnej ilości danych uczących oraz znaczny wzrost mocy obliczeniowej urządzeń dostępnych dla użytkowników. Algorytmy uczące sieci neuronowe są stale udoskonalane, a pewne teoretyczne ograniczenia sztucznych sieci okazały się w praktyce nieistotne.

Proste sieci neuronowe nie zawsze są wystarczające, szczególnie do zadań związanych z widzeniem komputerowym. Badania nad korą wzrokową w latach 60-tych doprowadziły do powstania wiele lat później koncepcji splotowej sieci neuronowej. Splotowe sieci neuronowe nie ograniczają się wyłącznie do postrzegania obrazów. Mają zastosowanie również w rozpoznawaniu mowy czy przetwarzaniu języka naturalnego. W tej pracy skupimy się głównie na zastosowaniach w przetwarzaniu obrazu.

Kluczowym rodzajem sieci neuronowych przedstawionym w tej pracy są autoenkodery. Stanowią one formę uczenia nienadzorowanego. Są szkolone po to, aby kopiować swoje dane wejściowe na wyjściu, przy zachowaniu pewnych ograniczeń, tak aby w procesie uczenia poznać istotne informacje o rozkładzie danych. Autoenkodery uczą się tzw. reprezentacji ukrytych, które zazwyczaj mają znacznie mniejszą wymiarowość niż dane wejściowe. Dzięki temu mogą służyć do redukowania wymiarowości, zwłaszcza w zadaniach wizualizacji. Mogą pełnić również funkcje wykrywaczy cech i służyć do wstępnego nienadzorowanego uczenia głębokich sieci neuronowych. Potrafią również generować nowe dane przypominające te ze zbioru uczącego.

W niniejszej pracy przedstawione zostaną pojęcia związane z sieciami neuronowymi, szczególnie splotowymi, oraz definicja autoenkodera, jego rodzaje i zastosowania. Następnie zostaną pokazane przykłady tych zastosowań.



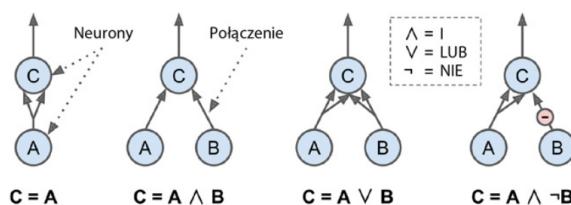
## Rozdział 1

### Podstawy teoretyczne

W tym rozdziale zostaną przedstawione pojęcia związane z autoenkoderami. Najpierw zostaną opisane ogólnie sztuczne sieci neuronowe, a także sieci splotowe stanowiące ważną część autoenkoderów. W następnej kolejności zostanie wprowadzona definicja autoenkodera, różne jego rodzaje pojawiające się w uczeniu nienadzorowanym, jak również przykłady ich zastosowań.

#### 1.1. Sztuczne sieci neuronowe

Sieci neuronowe są podziobrem uczenia maszynowego oraz stanowią kluczową część algorytmów głębokiego uczenia. Podstawowym elementem sztucznych sieci neuronowych jest sztuczny neuron, który do pewnego stopnia jest inspirowany budową biologicznego neurona. **Sztuczny neuron** jest systemem składającym się z co najmniej jednego binarnego wejścia i dokładnie jednego binarnego wyjścia. Wyjście uaktywnia się, jeżeli jest aktywna określona liczba wejść. Na rysunku 1.1 przedstawione są przykładowe sztuczne sieci neuronowe (SSN lub z angielskiego ANN) wykonujące różne operacje logiczne. W tych przykładach przyjęto, że neuron uaktywni się, gdy przynajmniej dwa wejścia będą aktywne [6].

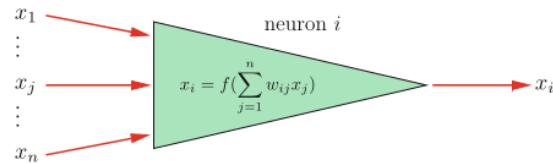


Rysunek 1.1: Przykładowe sztuczne sieci neuronowe rozwiązujące proste zadania logiczne

Źródło: [6]

Jedną z najprostszych architektur SSN jest **perceptron**, którego podstawą jest sztuczny neuron zwany **progową jednostką logiczną** (ang. *Threshold Logic Unit* - TLU) lub **liniową jednostką progową** (ang. *Linear Threshold Unit* - LTU). Wartościami wejść i wyjścia są liczby, a każde połączenie ma przyporządkowaną wagę. Jednostka TLU oblicza ważoną sumę

sygnałów wejściowych, a następnie zostaje użyta funkcja skokowa na tej sumie. Schemat takiej jednostki został przedstawiony na rysunku 1.2.



**Rysunek 1.2:** Struktura sztucznego neuronu, który stosuje funkcję skokową  $f$  na ważonej sumie sygnałów wejściowych

Źródło: [4]

Często używaną funkcją skokową jest **funkcja Heaviside'a**, określona równaniem

$$H(z) = \begin{cases} 0, & \text{jeśli } z < 0 \\ 1, & \text{jeśli } z \geq 0 \end{cases}$$

Czasami zamiast niej stosuje się również **funkcję signum**:

$$sgn(z) = \begin{cases} -1 & \text{jeśli } z < 0 \\ 0, & \text{jeśli } z = 0 \\ 1, & \text{jeśli } z > 0 \end{cases}$$

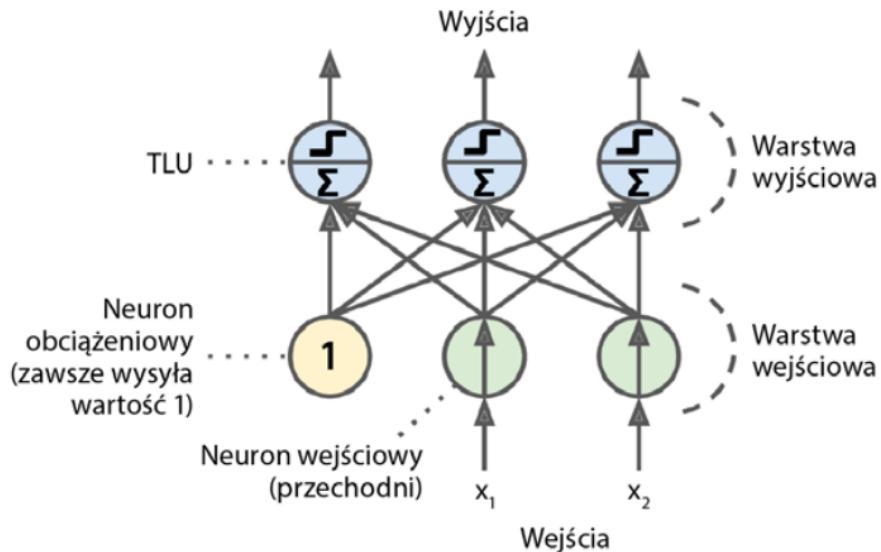
Perceptron jest złożony z jednej warstwy jednostek TLU, w której każdy neuron jest połączony ze wszystkimi wejściami. Warstwa tego typu nazywana jest **warstwą gęstą**. Warstwa, do której są dostarczane dane wejściowe, jest nazywana **warstwą wejściową** (ang. *input layer*). Najczęściej do tej warstwy jest wstawiany również **neuron obciążeniowy** (ang. *bias neuron*)  $x_0 = 1$ , który zawsze wysyła wartość 1. Na rysunku 1.3 znajduje się perceptron z dwoma neuronami wejściowymi i jednym obciążeniowym, a także z trzema neuronami w warstwie wyjściowej.

Obliczanie sygnałów wyjściowych w warstwie gęstej przedstawia się wzorem

$$h_{\mathbf{W}, \mathbf{b}}(\mathbf{X}) = \phi(\mathbf{X}\mathbf{W} + \mathbf{b})$$

gdzie  $\mathbf{X}$  - macierz cech wejściowych,  $\mathbf{W}$  - macierz wag połączeń (oprócz neuronu obciążeniowego),  $\mathbf{b}$  - wektor obciążzeń zawierający wagi połączeń neuronu obciążeniowego ze wszystkimi innymi neuronami,  $\phi$  - tzw. **funkcja aktywacji**, w przypadku TLU jest to funkcja skokowa.

Algorytm uczący, który służy do trenowania perceptronu, jest silnie inspirowany działaniem neuronu biologicznego. Gdy biologiczny neuron często pobudza inną komórkę nerwową, to połączenia między nimi stają się silniejsze. Reguła ta jest nazywana **regułą Hebb'a**.



**Rysunek 1.3:** Perceptron z trzema neuronami wejściowymi i trzema wyjściowymi

Źródło: [6]

Perceptrony są uczone za pomocą odmiany tej reguły, w której połączenia są wzmacniane, jeśli pomagają zmniejszyć wartość błędu. Dokładniej, w danym momencie perceptron przetwarza jeden przykład uczący i wylicza dla niego predykcję. Na każdy neuron wyjściowy odpowiadający za nieprawidłową prognozę następuje zwiększenie wag połączeń ze wszystkimi wejściami przyczyniającymi się do właściwej prognozy. Aktualizowanie wag przedstawia się następującym wzorem

$$\Delta w_{ij} = \eta (y_j - \hat{y}_j) x_i$$

gdzie  $w_{ij}$  - waga połączenia między  $i$ -tym neuronem wejściowym a  $j$ -tym neuronem wyjściowym,  $x_i$  -  $i$ -ta wartość wejściowa bieżącego przykładu uczącego,  $\hat{y}_j$  - wynik  $j$ -tego neuronu wyjściowego dla bieżącego przykładu uczącego,  $y_j$  - docelowy wynik  $j$ -tego neuronu,  $\eta$  - współczynnik uczenia.

Perceptron ma wiele wad związanych z niemożnością rozwiązywania pewnych trywialnych problemów (np. zadanie klasyfikacji rozłącznej czyli XOR). Część tych ograniczeń można wyeliminować, stosując architekturę SSN złożoną z wielu warstw perceptronów, czyli **perceptron wielowarstwowy** (ang. *Multi-Layer Perceptron*). Składa się on z jednej warstwy wejściowej (przechodniej), co najmniej jednej warstwy jednostek TLU - tzw. **warstwy ukryte** (ang. *latent layers*) i ostatniej warstwy jednostek TLU - warstwy wyjściowej. Oprócz warstwy wejściowej każda warstwa zawiera neuron obciążający i jest w pełni połączona z następną warstwą. Sieć zawierająca wiele warstw ukrytych nazywamy **głęboką siecią neuronową** (ang. *Deep Neural Network - DNN*).

Do uczenia perceptronów wielowarstwowych wykorzystywany jest algorytm **propagacji wstecznej** (ang. *backpropagation*). Propagacja wsteczna jest właściwie algorytmem

gradientu prostego [19]. Można go zapisać jako

$$w_{updated} = w_{old} - \eta \nabla E$$

gdzie  $E$  jest funkcją kosztu (funkcją straty) [19]. Proces jest powtarzany do momentu uzyskania zbieżności z rozwiązaniem, a każdy przebieg jest nazywany **epoką** (ang. *epoch*).

*Uwaga 1.1.* Wagi połączeń wszystkich warstw ukrytych należy koniecznie zainicjować losowo. W przeciwnym przypadku proces uczenia zakończy się niepowodzeniem. Na przykład jeśli wszystkie wagi i obciążenia zostaną zainicjowane wartością 0, to model będzie działał tak, jak gdyby składał się tylko z jednego neuronu. Przy zainicjowaniu wag losowo, symetria zostanie złamana i algorytm propagacji wstecznej będzie w stanie wytrenować zespół zróżnicowanych neuronów [6].

### 1.1.1. Funkcje aktywacji

Aby algorytm propagacji wstecznej działał prawidłowo, kluczową zmianą jest zastąpienie funkcji skokowej przez inne **funkcje aktywacji**. Zmiana ta jest konieczna, ponieważ funkcja skokowa zawiera jedynie płaskie segmenty i przez to nie pozwala korzystać z gradientu.

Najczęściej używana jest **funkcja logistyczna (sigmoidalna)**

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Ma ona w każdym punkcie zdefiniowaną pochodną niezerową, dzięki czemu algorytm gradientu prostego może na każdym etapie uzyskać lepsze wyniki. Zbiór wartości tej funkcji wynosi od 0 do 1.

Inną popularną funkcję aktywacji jest **tangens hiperboliczny**

$$\tanh(z) = 2\sigma(2z) - 1$$

Funkcja ta jest ciągła i różniczkowalna, a jej zakres wartości wynosi  $-1$  do  $1$ . Dzięki temu zakresowi wartości wynik każdej warstwy jest wyśrodkowany wobec zera na początku uczenia, co często pomaga w szybszym uzyskaniu zbieżności.

Wśród popularnych funkcji aktywacji należy także wyróżnić **funkcję ReLU** (ang. *Rectified Linear Unit* - prostowana jednostka liniowa) o wzorze

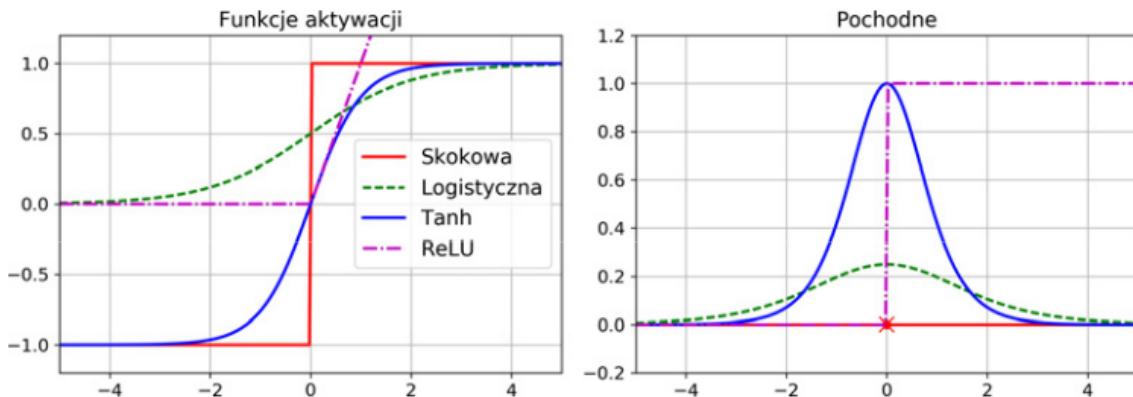
$$ReLU(z) = \max(0, z).$$

Jest ona ciągła, ale nieróżniczkowalna w punkcie 0. Jej pochodna dla  $z < 0$  wynosi zero. Jej atutem jest szybkość przetwarzania. Nie ma ona maksymalnej wartości wyjściowej. W

zadaniach regresji bywa wykorzystywany „wygładzony” wariant funkcji ReLU, czyli funkcja **softplus**:

$$\text{softplus}(z) = \log(1 + \exp(z))$$

Na rysunku 1.4 przedstawiono popularne funkcje aktywacji wraz z ich pochodnymi.



**Rysunek 1.4:** Przykładowe funkcje aktywacji wraz z pochodnymi

Źródło: [6]

### 1.1.2. Funkcje straty

Trenowanie sztucznej sieci neuronowej często polega na minimalizowaniu jakiejś funkcji straty. Opiszemy teraz kilka ważnych rodzajów takich funkcji:

1. Binarna entropia krzyżowa (ang. *binary cross-entropy*), używana do binarnej klasyfikacji, gdzie na wyjściu są możliwe dokładnie dwa stany[17]. Funkcja ta ma następującą postać

$$L = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)),$$

gdzie  $y_i$  jest oczekiwany wynikiem,  $\hat{y}_i$  wynikiem otrzymanym z modelu, a  $N$  jest liczbą obserwacji. W funkcji tej zawsze albo pierwszy albo drugi składnik jest równy zero (ponieważ  $y_i = 0$  lub  $y_i = 1$ ) Binarnej entropii krzyżowej można używać z funkcjami aktywacji takimi jak logistyczna (sigmoidalna), która w wyniku daje prawdopodobieństwo związane z wynikiem binarnym

2. Kategoryczna entropia krzyżowa (ang. *categorical cross-entropy*) jest stosowana w klasyfikacji wieloklasowej. Działa ona na tej samej zasadzie co entropia binarna, ale wyniki są sumowane dla więcej niż dwóch klas [17]. Jeżeli  $M$  jest liczbą klas, to entropia krzyżowa ma postać

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(\hat{y}_{ij})$$

3. Błąd średniokwadratowy (ang. *Mean Squared Error*) jest używany w zadaniach regresyjnych, gdzie oczekiwany wynikiem są liczby rzeczywiste [17]. Funkcja ta ma postać

$$L = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

W dalszej części pracy przy niektórych rozdziałach autoenkoderów opiszemy specjalne funkcje straty, które są w nich stosowane, jak na przykład dywergencja Kullbacka-Leiblera.

## 1.2. Sieci splotowe

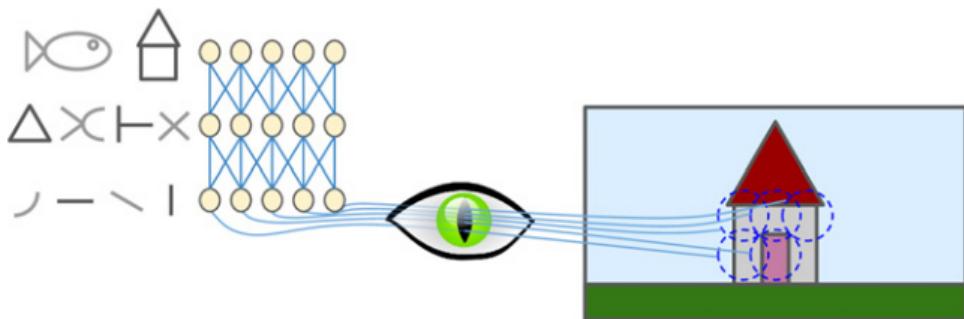
Autoenkodery, będące głównym tematem tej pracy, bardzo często składają się ze splotowych sieci neuronowych. W tym podrozdziale zostanie przybliżone pojęcie sieci splotowej oraz opisane zostaną jej elementy.

### 1.2.1. Czym jest sieć splotowa

**Splotowe sieci neuronowe** (ang. *convolutional neural networks*, CNN) są rodzajem sieci neuronowych służących do przetwarzania danych o znanej topologii siatki. Przykładem takich danych są szeregi czasowe, które można uznać za jednowymiarową siatkę z próbami w regularnych odstępach czasu, oraz dane graficzne, które można interpretować jako dwuwymiarową siatkę pikseli. Nazwa sieci splotowych pochodzi od wykorzystywanego przez te sieci działania matematycznego nazywanego **splotem** (konwolucją). Można powiedzieć, że sieci splotowe to po prostu sieci neuronowe, które w przynajmniej jednej z warstw zamiast ogólnego mnożenia macierzy wykorzystują splot [7]. Splotowe sieci neuronowe stanowią wynik badań nad korą wzrokową. Od lat 80-tych XX wieku są używane głównie rozpoznawania obrazów, w zagadnieniach takich jak klasyfikacja, detekcja obrazów, transfer stylu. Stanowią podstawę usług takich jak wyszukiwanie obrazu, inteligentne samochody, automatyczne systemy klasyfikowania filmów. Są skuteczne również w innych dziedzinach niż jedynie komputerowe widzenie - takie dziedziny to na przykład rozpoznawanie mowy (*voice recognition*) oraz przetwarzanie języka naturalnego (*Natural Language Processing, NLP*) [6].

Ponieważ splotowe sieci neuronowe są silnie inspirowane działaniem biologicznej kory wzrokowej, to ten podrozdział zostanie poświęcony wyjaśnieniu jej działania. Na podstawie badań prowadzonych pod koniec lat 50-tych XX wieku [8] [9] wykazano, że neurony biologiczne w korze wzrokowej reagują na określone wzorce w niewielkich obszarach pola wzrokowego, zwanych **polami receptivejnymi**, a w miarę przepływu sygnału wzrokowego przez kolejne moduły w mózgu neurony rozpoznają coraz bardziej skomplikowane wzorce wykrywane w coraz większych polach receptivejnych. Wiele neuronów stanowiących korę

wzrokową tworzy **lokalne pola recepcyjne**, reagujące jedynie na bodźce wzrokowe mieszczące się w określonym rejonie pola wzrokowego, przy czym lokalne pola recepcyjne poszczególnych neuronów mogą się na siebie nakładać. Takie pola recepcyjne łącznie tworzą całe pole wzrokowe. Dodatkowo badacze zauważyli, iż pewne neurony reagują wyłącznie na obrazy składające się z linii poziomych, lub innych linii ułożonych w konkretny sposób (dwa neurony mogą nawet mieć to samo pole recepcyjne, ale reagować na różne ułożenie linii). Badacze stwierdzili, że niektóre komórki nerwowe mają większe pola recepcyjne i wykrywają bardziej skomplikowane kształty. Takie neurony, odpowiedzialne za rozpoznawanie bardziej skomplikowanych kształtów, znajdują się na wyjściu neuronów reagujących na prostsze bodźce. Działanie neuronów biologicznych w korze wzrokowej, zgodnie z powyższym opisem, zostało zobrazowane na rysunku 1.5.



**Rysunek 1.5:** Działanie neuronów biologicznych w korze wzrokowej

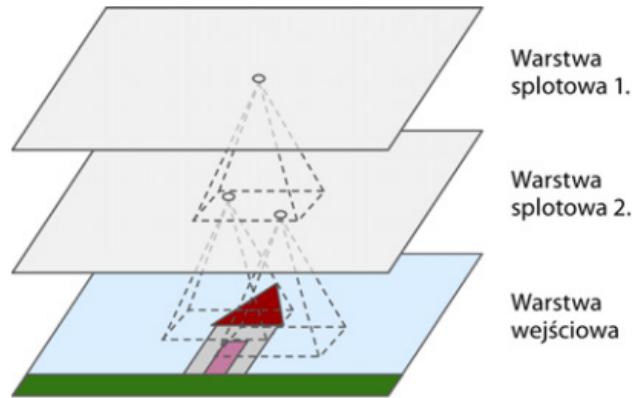
Źródło: [6]

### 1.2.2. Elementy sieci splotowej

#### Warstwy splotowe

Najistotniejszą częścią sieci CNN jest **warstwa splotowa** (*convolutional layer*). Na rysunku 1.6 widoczny jest przykład warstw splotowych z prostokątnymi polami recepcyjnymi. W pierwszej warstwie splotowej neurony nie są połączone z każdym pikselem obrazu wejściowego (w przeciwieństwie do opisanych wcześniej warstw gęstych), lecz jedynie z pikselami znajdującymi się w polu recepcyjnym danego neuronu. W kolejnej warstwie każdy neuron łączy się wyłącznie z neuronami z niewielkiego obszaru pierwszej warstwy. Dzięki temu sieć koncentruje się na pewnych cechach w pierwszej warstwie, a w drugiej warstwie może je łączyć w bardziej złożone kształty. Taka hierarchiczna struktura w naturalny sposób występuje na zdjęciach, co przyczyniło się do dużej skuteczności sieci splotowych w rozpoznawaniu obrazu.

Neuron znajdujący się w wierszu  $i$  oraz kolumnie  $j$  danej warstwy jest połączony z wyjściami neuronów poprzedniej warstwy zlokalizowanymi w rzędach od  $i$  do  $i + f_h - 1$  i

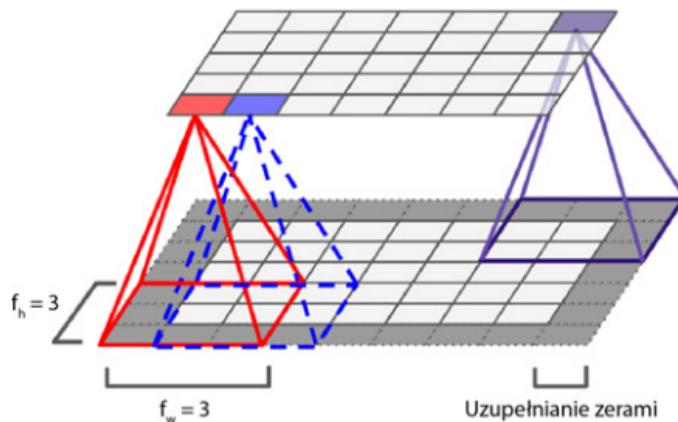


**Rysunek 1.6:** Warstwy splotowe z prostokątnymi lokalnymi polami recepcyjnymi

Źródło: [6]

kolumnach od  $j$  do  $j + f_w - 1$ , gdzie  $f_h$  i  $f_w$  oznaczają, odpowiednio, wysokość i szerokość pola recepcyjnego (rysunek 1.7). W celu uzyskania takich samych wymiarów każdej warstwy najczęściej są dodawane zera wokół wejść, co zostało pokazane na rysunku 1.7. Proces ten nazywamy **uzupełnianiem zerami** (ang. zero padding).

*Uwaga 1.2 (Dopełnianie).* Dopełnianie (padding) jest ściśle związane z parametrem kroku (który zostanie opisany w kolejnym akapicie). Zapewnia poprawność obliczeń w warstwie konwolucyjnej [10].

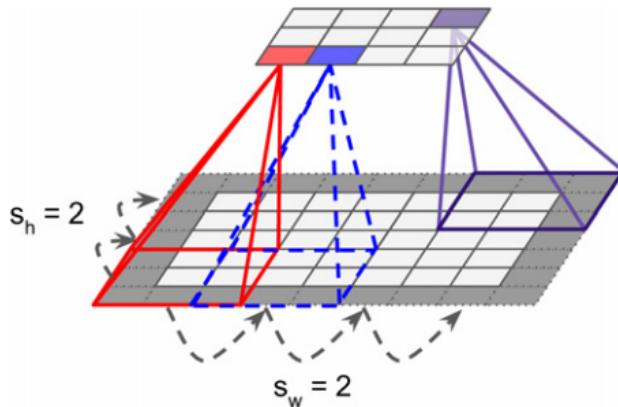


**Rysunek 1.7:** Związek pomiędzy warstwami a uzupełnianiem zerami

Źródło: [6]

Możliwe jest również łączenie bardzo dużej warstwy wejściowej ze znacznie mniejszą kolejną warstwą poprzez rozdzielenie pól recepcyjnych, tak jak zaprezentowano na rysunku 1.8. Rozwiążanie to zmniejsza drastycznie złożoność obliczeniową modelu. Odległość pomiędzy dwoma kolejnymi polami recepcyjnymi nosi nazwę **kroku** (ang. stride). Na widocznym schemacie warstwa wejściowa o wymiarach  $5 \times 7$  (plus uzupełnianie zerami) łączy się z warstwą o rozmiarze  $3 \times 4$  za pomocą pól recepcyjnych będących kwadratami  $3 \times 3$  i

kroku o wartości 2 (w omawianym przykładzie krok jest taki sam w obydwu wymiarach, ale nie jest to wcale regułą). Neuron zlokalizowany w rzędzie  $i$  oraz kolumnie  $j$  górnej warstwy łączy się z wyjściami neuronów dolnej warstwy mieszczącymi się w rzędach od  $i \times s_h$  do  $i \times s_h + f_h - 1$  i w kolumnach od  $j \times s_w$  do  $j \times s_w + f_w - 1$ , gdzie  $s_h$  i  $s_w$  definiują wartości kroków odpowiednio w kolumnach i rzędach.



Rysunek 1.8: Warstwa splotowa z krokiem o długości 2

Źródło: [6]

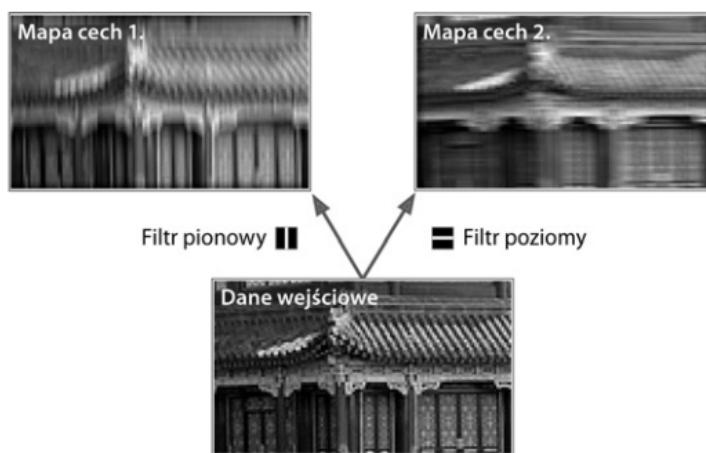
*Uwaga 1.3 (Długość kroku).* Długość kroku (stride length) - oznacza odległość, o jaką jądro przesuwa się po obrazie. Często stosowaną wielkością jest długość jednego piksela, również dwa piksele, a rzadziej trzy. Dłuższe kroki nie są stosowane, ponieważ jądro może wtedy pomijać obszary obrazu potencjalnie wartościowe dla modelu. Z drugiej strony, im dłuższy krok, tym większa szybkość uczenia się modelu, ponieważ jest mniej obliczeń do wykonania. Trzeba znajdować kompromis między tymi efektami [10].

## Filtry

Wagi neuronu mogą być przedstawiane jako niewielki obraz o rozmiarze pola receptivejnego. Na przykład na rysunku 1.9 widoczne są dwa możliwe zbiory wag, tak zwane **filtre** (lub **jadra splotowe**; ang. *convolution kernels*). Pierwszy filtr jest symbolizowany jako czarny kwadrat z białą pionową linią przechodzącą przez jego środek (jest to macierz o wymiarach  $7 \times 7$  wypełniona zerami oprócz środkowej kolumny, która zawiera jedynki); neurony zawierające te wagi będą ignorować wszystkie elementy w polu receptivejnym oprócz znajdujących się w środkowej pionowej linii (dane wejściowe znajdujące się poza tą linią będą przemnażane przez 0). Drugi filtr wygląda podobnie. Różnica polega na tym, że środkowa linia jest ułożona poziomo. Także w tym wypadku będą brane pod uwagę jedynie dane wejściowe znajdujące się w tej linii.

Jeśli wszystkie neurony w danej warstwie będą korzystać z tego samego filtra „pionowego” (i takiego samego członu obciążenia), a do sieci wczytamy obraz zaprezentowany

na dole rysunku 1.9, to uzyskamy obraz widoczny w lewym górnym rogu rysunku. Można zauważyc, że po zastosowaniu tego filtru pionowe białe linie stają się wyraźniej widoczne, natomiast pozostała część obrazu zostaje rozmazana. W analogiczny sposób otrzymujemy obraz widoczny w prawym górnym rogu rysunku po zastosowaniu filtru „poziomego”; w tym przypadku białe poziome linie zostają wyostrzone, a reszta obrazu ulega zamazaniu. Zatem warstwa wypełniona neuronami wykorzystującymi ten sam filtr daje nam **mapę cech** (ang. *feature map*), dzięki której możemy dostrzec elementy najbardziej przypominające dany filtr. Sieć CNN w czasie uczenia wyszukuje filtry najbardziej przydatne do danego zadania i uczy się łączyć je w bardziej złożone wzorce [6].



**Rysunek 1.9:** Uzyskiwanie dwóch map cech za pomocą dwóch różnych filtrów

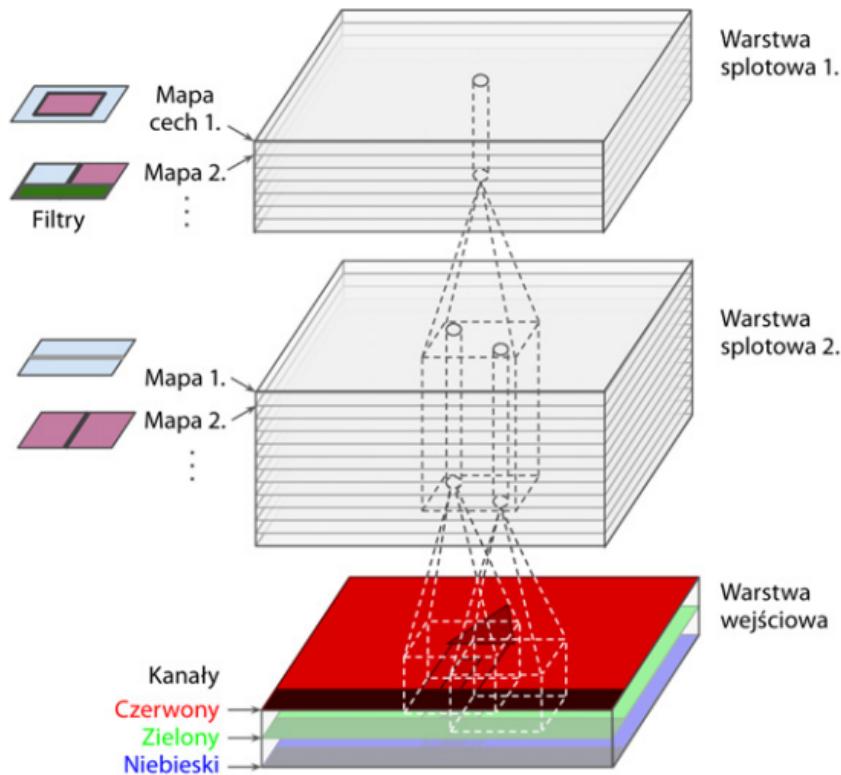
Źródło: [6]

*Uwaga 1.4 (Wielkość jądra).* Jądro (zwane również filtrem lub polem receptywnym) typowo ma wysokość i szerokość trzech pikseli. Rozmiar ten okazał się optymalny w szerokim zakresie zastosowań widzenia maszynowego w nowoczesnych sieciach konwolucyjnych. Popularna jest również wielkość 5x5 pikseli, a maksymalny stosowany rozmiar to 7x7 pikseli. Jeśli jądro jest zbyt duże w stosunku do obrazu, wtedy w polu receptywnym pojawia się zbyt wiele cech i warstwa konwolucyjna nie jest w stanie się skutecznie uczyć. Jeżeli jądro jest zbyt małe, np. ma wymiary 2x2 piksele, nie jest w stanie dopasować się do żadnej struktury, przez co jest bezużyteczne [10].

### Stosy map cech

Warstwa splotowa była do tej pory dla uproszczenia przedstawiana w postaci dwuwy- miarowej warstwy, ale w rzeczywistości składa się ona z kilku map cech o identycznych roz- miarach, dlatego trójwymiarowe odwzorowanie jest bliższe rzeczywistości (rysunek 1.10). W zakresie jednej mapy cech każdy neuron jest przydzielony do jednego piksela, a wszystkie tworzące ją neurony współdzielą te same parametry (wagi i człon obciążenia). Neurony w

innych mapach cech mają odmienne wartości parametrów. Pole recepcyjne neuronu nie ulega zmianie, ale „przebiega” przez wszystkie mapy cech poprzednich warstw.



**Rysunek 1.10:** Warstwy splotowe zawierające wiele map cech, a także zdjęcie z trzema kanałami barw

źródło: [6]

Co więcej, obrazy wejściowe także składają się z kilku warstw podrzędnych, po jednej na każdy **kanał barw** (ang. *color channel*). Standardowo występują trzy kanały barw — czerwony, zielony i niebieski (ang. red, green, blue — RGB). Obrazy czarno-białe (w odcieniach szarości) zawierają tylko jeden kanał, ale istnieją też takie zdjęcia, które mogą mieć ich znacznie więcej — np. fotografie satelitarne utrwalające dodatkowe częstotliwości fal elektromagnetycznych (takie jak podczerwień).

W szczególności neuron zlokalizowany w rzędzie  $i$  oraz kolumnie  $j$  mapy cech  $k$  w danej warstwie splotowej  $l$  jest połączony z neuronami wcześniejszej warstwy  $l - 1$  umieszczonymi w rzędach od  $i \times s_h$  do  $i \times s_h + f_h - 1$  i kolumnach od  $j \times s_w$  do  $j \times s_w + f_w - 1$  we wszystkich mapach cech (warstwy  $l - 1$ ). Wszystkie neurony znajdujące się w tym samym rzędzie  $i$  oraz kolumnie  $j$ , ale w innych mapach cech są połączone z wyjściami dokładnie tych samych neuronów poprzedniej warstwy.

Powyższy opis został podsumowany następującym wzorem, służącym do obliczania wyniku danego neuronu w warstwie splotowej:

$$z_{i,j,k} = b_k \sum_{u=0}^{f_h-1} \sum_{v=0}^{f_w-1} \sum_{k'=0}^{f_{n'}-1} x_{i',j',k'} \cdot w_{u,v,k',k}, \quad \text{gdzie } \begin{cases} i' = i \times s_h + u \\ j' = j \times s_w + v \end{cases}$$

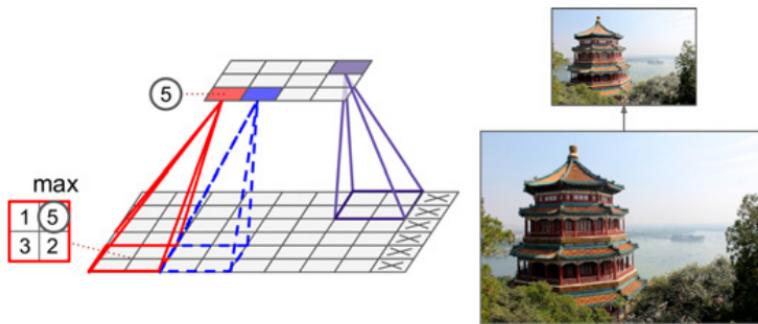
W tym równaniu:

- $z_{i,j,k}$  jest wyjściem neuronu znajdującego się w rzędzie  $i$ , kolumnie  $j$  i mapie cech  $k$  warstwy splotowej  $l$ ;
- jak już zostało wyjaśnione,  $s_h$  i  $s_w$  to kroki pionowy i poziomy,  $f_h$  i  $f_w$  są wysokością i szerokością pola recepcyjnego, natomiast  $f_{n'}$  oznacza liczbę map cech w poprzedniej warstwie ( $l - 1$ );
- $x_{i',j',k'}$  jest wyjściem neuronu zlokalizowanego w warstwie  $l - 1$ , rzędzie  $i'$ , kolumnie  $j'$ , mapie cech  $k$  (lub kanale  $k'$ , jeżeli poprzednia warstwa była warstwą wejściową);
- $b_k$  to człon obciążenia dla mapy cech  $k$  (w warstwie  $l$ ); można go interpretować jako „pokrętło jasności” mapy cech  $k$ ;
- $w_{u,v,k',k}$  jest wagą połączenia pomiędzy dowolnym neuronem w mapie cech  $k$  warstwy  $l$  a jego wejściem mieszącym się w wierszu  $u$ , kolumnie  $v$  (względem pola recepcyjnego neuronu) a mapą cech  $k'$ .

### Warstwa łącząca

Oprócz warstw splotowych, ważnym elementem budulcowym sieci CNN jest **warstwa łącząca**, zwana także redukującą (ang. pooling layer). Warstwa konwolucyjna może zawierać dowolną liczbę jąder, z których każde generuje mapę aktywacji. Zatem wyjściem warstwy konwolucyjnej jest trójwymiarowa tablica aktywacji, której głębokość jest równa liczbie filtrów. Warstwa redukująca zmniejsza przestrzenny wymiar mapy aktywacji, pozostawiając jej głębokość bez zmian [10]. Jej celem jest podpróbkowanie (ang. subsample; tj. zmniejszenie) obrazu wejściowego w celu zredukowania obciążenia obliczeniowego, wykorzystania pamięci i liczby parametrów (a tym samym ograniczenia ryzyka przetrenowania). Podobnie jak w przypadku warstw splotowych, każdy neuron stanowiący część warstwy łączącej łączy się z wyjściami określonej liczby neuronów warstwy poprzedniej, mieszącymi się w obszarze niewielkiego, prostokątnego pola recepcyjnego. Podobnie jak wcześniej, musimy definiować rozmiar tego pola, wartość kroku, rodzaj uzupełniania zerami itd. Jednakże warstwa łącząca nie zawiera żadnych wag; jej jedynym zadaniem jest gromadzenie danych wejściowych za pomocą jakiejś funkcji agregacyjnej, np. maksymalizującej lub średniającej. Na rysunku 1.11 przedstawiony jest najpopularniejszy rodzaj warstwy łączącej — **maksymalizująca warstwa łącząca** (ang. max pooling layer). W tym przykładzie korzystamy z **jądra łączącego** (ang. pooling kernel) o rozmiarze  $2 \times 2$ , kroku o wartości 2 i z pominięciem uzupełniania

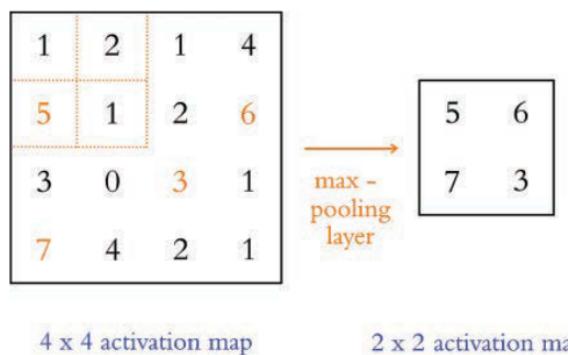
zerami. Jedynie maksymalna wartość z każdego jądra zostaje przekazana do następnej warstwy, natomiast pozostałe wartości wejściowe zostają odrzucone. Na przykład w lewym dolnym polu recepcyjnym na rysunku 1.11 widzimy wartości wejściowe 1, 5, 3, 2, zatem tylko wartość maksymalna, czyli 5, zostanie przekazana do następnej warstwy. Z powodu kroku równego 2 obraz wyjściowy ma szerokość i wysokość o połowę mniejsze w porównaniu do obrazu wejściowego (zaokrąglamy w dół, ponieważ nie korzystamy z uzupełniania zerami).



**Rysunek 1.11:** Maksymalizująca warstwa łącząca (jądro łączące:  $2 \times 2$ , krok: 2, brak uzupełniania zerami)

Źródło: [6]

*Uwaga 1.5* (Parametry warstwy redukującej). Filtr w warstwie redukującej ma zazwyczaj wymiary  $2 \times 2$  piksele, a krok ma długość dwóch pikseli. W takim wypadku filtr w każdej pozycji przetwarza cztery wartości aktywacji, wybiera największą i w efekcie czterokrotnie redukuje liczbę aktywacji [10].



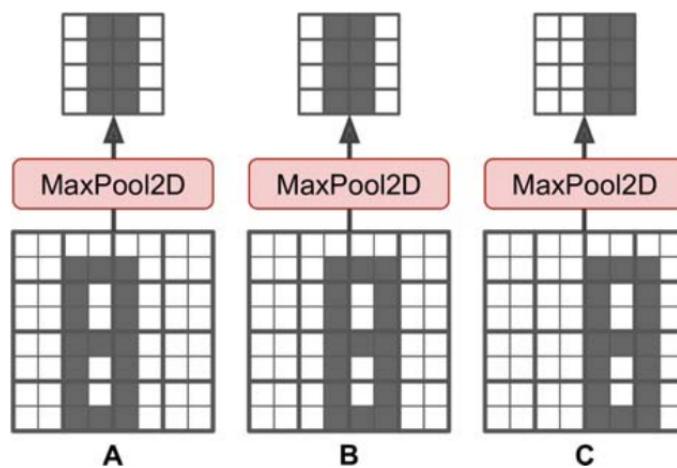
**Rysunek 1.12:** Warstwa *max-pooling* z filtrem i krokiem rozmiaru  $2 \times 2$ , zastosowana do mapy aktywacji o rozmiarze  $4 \times 4$  (widoczna po lewej stronie)

Źródło: [10]

Na rysunku 1.12 widoczne jest działanie maksymalizującej warstwy redukującej na mapie aktywacji o wymiarze  $4 \times 4$ . Filtr przesuwa się nad danymi wejściowymi od lewej do prawej, od góry do dołu, tak jak w warstwie konwolucyjnej, i w każdej zajmowanej pozycji

przeprowadza operację redukcji danych. W tym przykładzie filtr i krok mają rozmiar 2x2. Skutkuje to otrzymaniem mapy cztery razy mniejszej niż oryginalna.

Oprócz ograniczania liczby obliczeń, zużycia pamięci i liczby parametrów maksymalizująca warstwa łącząca wprowadza także pewien stopień **niezmienniczości** w stosunku do drobnych przesunięć, co widać na rysunku 1.13. Zakładamy w tym przykładzie, że piksele jasne mają mniejszą wartość od pikseli ciemnych. Trzy obrazy (A, B i C) przechodzą przez maksymalizującą warstwę łączącą o jądrze  $2 \times 2$  i kroku równym 2. Obrazy B i C wyglądają tak samo jak obraz A, ale są przesunięte o odpowiednio jeden i dwa piksele w prawo. Jak widać, rezultaty wygenerowane w maksymalizującej warstwie łączącej z obrazów A i B są identyczne. Na tym polega **niezmienniczość przesunięć** (ang. translation invariance). W przypadku obrazu C wynik jest odmienny: jest on przesunięty o jeden piksel w prawo (nadal jednak pozostaje niezmieniony w mniej więcej 75%). Poprzez wstawianie maksymalizującej warstwy łączącej co kilka warstw sieci CNN możliwe jest uzyskanie ograniczonej niezmienniczości przesunięć w większej skali. Ponadto warstwa ta zapewnia niewielki stopień niezmienniczości rotacyjnej i drobną niezmienniczość skalowania. Tego typu niezmienniczość (mimo że jest ograniczona) jest przydatna w zagadnieniach, w których prognozy nie powinny być zależne od tych zmian, na przykład w zadaniach klasyfikacji.



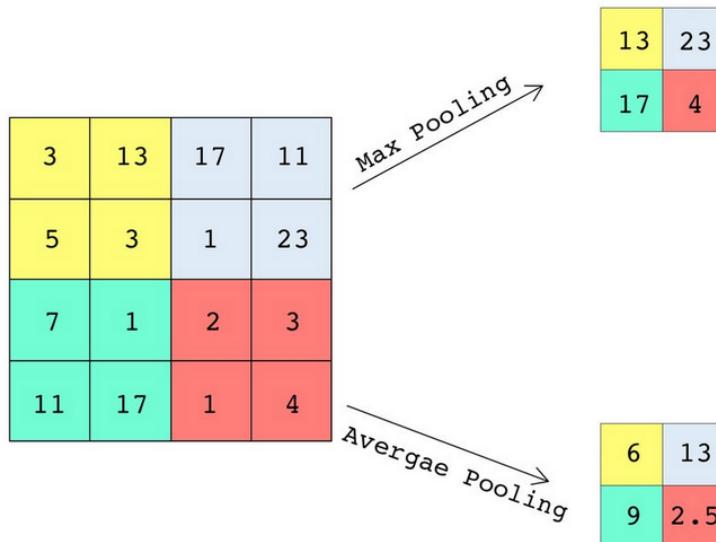
**Rysunek 1.13:** Niezmienniczość związana z drobnymi przesunięciami

Źródło: [6]

Maksymalizująca warstwa łącząca ma również pewne wady. Przede wszystkim jest ona bardzo destrukcyjna: nawet w przypadku niewielkiego jądra o rozmiarze  $2 \times 2$  i kroku o wartości 2 dane wyjściowe będą dwukrotnie mniejsze w każdym kierunku (zatem obszar obrazu będzie zmniejszony czterokrotnie), co oznacza porzucenie 75% wartości wejściowych. Z kolei w pewnych zastosowaniach niezmienniczość jest niepożądana, na przykład w segmentacji semantycznej (zadaniu klasyfikowania każdego piksela obrazu zgodnie z jego przynależnością do danego obiektu): jest oczywiste, że jeżeli obraz wejściowy zostanie

przesunięty o jeden piksel w prawo, to wynik również powinien być przesunięty w taki sam sposób. Wówczas celem staje się **ekwiwariancja** (ang. equivariance), a nie niezmienniczość: mała zmiana w sygnale wejściowym powinna prowadzić do powiązanej z nią niewielkiej zmiany w sygnale wyjściowym.

W sposób analogiczny do maksymalizującej warstwy łączącej jest zdefiniowana **uśredniająca warstwa łącząca** (average pooling layer), która zamiast maksimum używa średniej. Jest ona rzadziej wybierana w zastosowaniach niż warstwa maksymalizująca, z powodu słabszej wydajności. Obliczanie średniej zazwyczaj powoduje mniejszą utratę informacji niż obliczanie maksimum, ale za to warstwa maksymalizująca zachowuje wyłącznie najistotniejsze cechy i ignoruje te mniej ważne, dlatego kolejne warstwy otrzymują coraz czystszy sygnał. Na rysunku 1.14 widoczne jest porównanie maksymalizującej i uśredniającej warstwy redukujących.



Rysunek 1.14: Porównanie warstwy łączącej maksymalizującej i uśredniającej

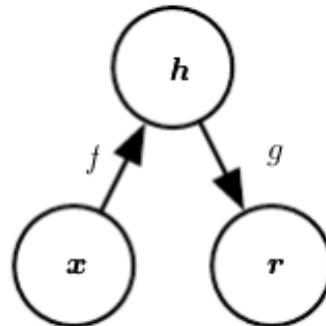
Źródło: [6]

Ostatnim rodzajem warstwy łączącej często spotykanym we współczesnych architekturach, jest **globalna uśredniająca warstwa łącząca** (ang. global average pooling layer). Mechanizm jej działania jest całkiem odmienny: oblicza ona jedynie średnią każdej mapy cech (przypomina to działanie uśredniającej warstwy łączącej, w której jądro ma takie same wymiary przestrzenne jak dane wejściowe). Oznacza to, że generuje ona na wyjściu pojedynczą wartość na każdą mapę cech i na każdy przykład. Jest to rozwiązanie skrajnie destrukcyjne (większość informacji zawartych w mapie cech zostaje utraconych), ale bywa przydatne na wyjściu modelu [6].

## 1.3. Autoenkodery

### 1.3.1. Czym jest autoenkoder

**Autoenkoder** (nazywany także autokoderem, z ang. *autoencoder*, *auto-encoder*) jest rodzajem sieci neuronowej przeznaczonym głównie do kodowania danych wejściowych do skompresowanej i znaczącej reprezentacji, a następnie dekodowania ich z powrotem w taki sposób, aby zrekonstruowane dane były jak najbardziej podobne do oryginalnych [3]. Autoenkodery uczą się gęstych reprezentacji danych, tzw. **reprezentacji ukrytych** (ang. *latent representations*) lub **kodowań** (ang. *codings*) w formie uczenia nienadzorowanego. Kodowania często mają mniejszą wymiarowość od danych wejściowych, dzięki czemu autoenkodery mogą służyć do redukcji wymiarowości. Mają też zastosowanie w modelach generatywnych (ang. *generative models*), które potrafią losowo generować nowe dane przypominające zbiór uczący, choć warto zaznaczyć, że często lepszej jakości dane można uzyskać przy użyciu generatywnych sieci przeciwnostawnych, czyli GAN (ang. *Generative Adversarial Networks*). [6]. Autoenkoder jest zatem siecią neuronową szkoloną po to, aby kopiować dane wejściowe do wyjścia. Zawiera ukrytą warstwę  $h$ , która opisuje kodowanie używane do reprezentowania wejścia. Sieć można postrzegać jako składającą się z dwóch części: kodującej funkcji  $h(x)$  i dekodera, który tworzy rekonstrukcję  $r = g(h)$  [7]. Ogólna struktura autoenkodera jest przedstawiona na rysunku 1.15.



**Rysunek 1.15:** Struktura autoenkodera odwzorowującego wejście  $x$  na wyjście  $r$  (nazywane rekonstrukcją) poprzez reprezentację ukrytą (kodowanie)  $h$ . Autoenkoder składa się z dwóch składników: kodera  $f$  (odwzorowującego  $x$  na  $h$ ) i dekodera  $g$  (odwzorowującego  $h$  na  $r$ )

Źródło: [7]

Gdyby autoenkoder nauczył się po prostu, aby na wyjściu ustawać zawsze  $g(f(x)) = x$ , to nie byłby zbyt przydatny. Z tego powodu autoenkodery są projektowane tak, aby nie potrafiły kopiować w sposób doskonały. Zazwyczaj nakładane są ograniczenia, aby autoenkoder mógł kopiować jedynie w przybliżeniu, i tylko takie dane, które są podobne do danych ze

zbioru uczącego. Dzięki temu model musi wybierać jedynie pewne aspekty danych wejściowych, które powinny być kopiowane. W ten sposób może on zdobyć użyteczne informacje o strukturze danych.

Autoenkodery można wyobrazić sobie jako specjalny przypadek sieci jednokierunkowych i można je szkolić, używając wszystkich tych samych technik, zwykle minipakietowego spadku gradientu po gradientach obliczonych przez propagację wstecz. W przeciwieństwie do ogólnych sieci jednokierunkowych, autoenkodery można szkolić za pomocą recyrkulacji, czyli algorytmu uczącego się na bazie porównywania aktywacji sieci na oryginalnym wejściu z aktywacjami na zrekonstruowanym wejściu [7].

### 1.3.2. Rodzaje autoenkoderów

Jak wspomniano wcześniej, aby autoenkoder był użyteczny, zamiast jedynie kopiować wejście do wyjścia, można na niego nałożyć różne ograniczenia, jak na przykład limit na rozmiar reprezentacji ukrytej. Ze względu na nakładane ograniczenia, wyróżniamy wiele rodzajów autoenkoderów, a wśród nich:

- autoenkodery niedopełnione (ang. *undercomplete*), w których wyjście musi mieć mniejszy wymiar niż wejście
- autoenkodery z regularyzacją (ang. *regularized*), w których wyjście ma taki sam lub większy (ang. *overcomplete*) wymiar niż wyjście, ale używają specjalnie dopasowanych funkcji straty. Wśród autoenkoderów z regularyzacją można rozróżnić na przykład:
  - autoenkodery rzadkie (ang. *sparse*), które dążą do rzadkiej reprezentacji ukrytej
  - autoenkodery odszumiające (ang. *denoising*), które na wejściu dostają zniekształcone dane, a starają się odzyskać pierwotne, niezaszumione informacje
  - autoenkodery kurczliwe (ang. *contractive*) dążące do małego rozmiaru pochodnej
- autoenkodery stosowe (ang. *stacked*) nazywane również głębokimi (ang. *deep*)
- autoenkodery splotowe (ang. *convolutional*)
- autoenkodery rekurencyjne (ang. *recurrent*)
- autoenkodery wariancyjne (ang. *variational*)
- autoenkodery przeciwnostawne (ang. *adversarial*)

W następnych podrozdziałach zostaną przybliżone cechy charakterystyczne tych rodzajów autoenkoderów.

#### Autoenkodery niedopełnione

Kopiowanie wejścia do wyjścia może wydawać się bezużyteczne, ale wyjście dekodera niekoniecznie jest głównym celem. Często najistotniejszym skutkiem przeszkolenia autoenkodera do kopiowania będzie kodowanie ukryte  $h$ , mające przydatne właściwości. Jednym ze sposobów, aby uzyskać przydatne cechy z autoenkodera, jest ograniczenie  $h$  do mniejszego

wymiaru niż  $x$ . Autoenkoder, w którym wymiar kodu jest mniejszy niż wymiar wejściowy, jest nazywany niekompletnym (niedopełnionym). Poznawanie niekompletnych reprezentacji zmusza autoenkoder do przechwycenia najistotniejszych cech danych szkoleniowych. Proces poznawania jest opisywany jako minimalizowanie funkcji straty

$$L(x, g(f(x)))$$

gdzie  $L$  jest funkcją straty karzącą  $g(f(x))$  za niepodobieństwo do  $x$  jak np. błąd średniokwadratowy.

Gdy dekoder jest liniowy, a  $L$  to błąd średniokwadratowy, niekompletny autoenkoder uczy się obejmować tą samą podprzestrzeń co PCA. W tym przypadku poznanie zasadniczej podprzestrzeni danych szkoleniowych przez szkolony do kopiowania autoenkoder jest efektem ubocznym. Autoenkodery z nieliniowymi funkcjami kodowania  $f$  i nieliniowymi funkcjami dekodowania  $g$  mogą więc poznawać potężniejsze, nieliniowe uogólnienie PCA.

Jeśli jednak koder i dekoder będą mieć zbyt dużą pojemność, autoenkoder może nauczyć się wykonywać kopiowanie bez wyodrębniania pożytecznych informacji o rozkładzie danych. Teoretycznie można sobie wyobrazić, że autoenkoder z jednowymiarowym kodem, ale bardzo potężnym nieliniowym koderem, może nauczyć się reprezentować każdy przykład szkoleniowy  $x^{(i)}$  za pomocą kodu  $i$ . Dekoder mógłby nauczyć się odwzorowywać te całkowitoliczbowe indeksy z powrotem na wartości konkretnych przykładów szkoleniowych. Ten konkretny przykład nie występuje w praktyce, ale pokazuje wyraźnie, że autoenkoder przeszkolony do wykonywania kopiowania może zawieść, jeśli chodzi o poznanie czegoś przydatnego na temat zbioru danych, jeśli pozwoli się, aby miał za dużą pojemność [7].

### Autoenkodery z regularyzacją

W przypadku, gdy wymiar wyjścia jest większy (autoenkodery nadkompletne) lub równy niż wymiar wejścia, nawet liniowy koder i dekoder mogą nauczyć się kopiować wejście do wyjścia bez uczenia się niczego przydatnego na temat rozkładu danych. Ideałem byłaby możliwość udanego szkolenia dowolnej architektury autoenkodera przy wyborze wymiaru kodu oraz pojemności kodera i dekodera na podstawie złożoności rozkładu modelowanego. Można to zrobić, stosując regularyzację. Zamiast ograniczać pojemność modelu przez zachowywanie płytkości kodera i dekodera oraz małego rozmiaru kodu, autoenkodery z regularyzacją używają funkcji straty, dzięki której model może posiadać inne właściwości oprócz możliwości kopiowania swojego wejścia do wyjścia. Do tych właściwości należą:

- rzadkość reprezentacji
- mały rozmiar pochodnej reprezentacji
- odporność na szum lub brakujące dane wejściowe

Autoenkoder z regularizacją może być nieliniowy i nadkompletny, a mimo to nauczyć się czegoś wartościowego o rozkładzie danych, nawet jeśli pojemność modelu jest na tyle duża, aby poznać trywialną funkcję tożsamościową [7].

### Rzadkie autoenkodery

Rzadkie autoenkodery są zwykle używane do tego, aby uczyć się cech do innego zadania, takiego jak klasyfikacja. Autoenkoder, który dzięki regularizacji jest rzadki, musi reagować na unikatowe statystyczne cechy zbioru danych, na którym został wyszkolony, a nie tylko działać jak funkcja tożsamościowa.

Kryterium szkolenia autoenkodera rzadkiego obejmuje karę rzadkości  $\Omega(h)$  na warstwie kodu  $h$  oprócz błędu rekonstrukcji

$$L(x, g(f(x))) + \Omega(h)$$

gdzie  $g(h)$  to wyjście dekodera, a zwykle mamy  $h = f(x)$ , czyli wyjście kodera [7]. Dodanie tego składnika do funkcji kosztu zmusza autoenkoder do zmniejszenia liczby aktywnych neuronów w warstwie kodowania. W ten sposób każde wejście musi być reprezentowane jako kombinacja niewielkiej liczby pobudzeń. Dzięki temu każdy neuron warstwy kodowania zazwyczaj uczy się wykrywać jakąś przydatną cechę [6].

W każdym przebiegu uczenia następuje pomiar rzeczywistej rzadkości warstwy kodowania i karanie modelu, gdy zmierzona rzadkość różni się od docelowej. W tym celu obliczania jest średnia aktywacja każdego neuronu w warstwie dla całej grupy przykładów uczących. Rozmiar tej grupy nie może być zbyt mały, aby wyliczona wartość średniej była dokładna. Następnie nakładana jest kara na zbyt aktywne neurony, poprzez dodanie funkcji straty rzadkości (ang. *sparsity loss*)  $\Omega(h)$  do funkcji kosztu. Dla przykładu, jeśli średnia wartość aktywacji neuronu to 0.3, ale docelowo powinna wynosić 0.1, musimy ją zmniejszyć. Jednym ze sposobów jest dodanie kwadratu błędu  $(0.3 - 0.1)^2$  do funkcji kosztu. Lepszym rozwiązaniem w praktyce jest zastosowanie dywergencji Kullbacka-Leiblera, której gradien-

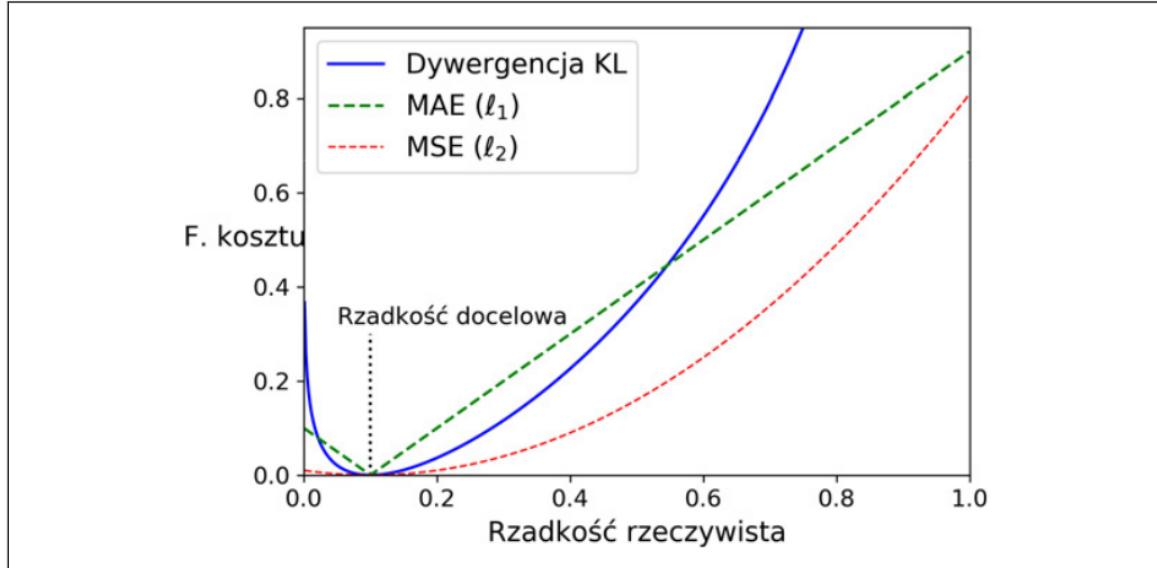
ty są znacznie większe niż w błędzie średniokwadratowym (rysunek 1.16).

Mając dwa dyskretne rozkłady prawdopodobieństwa  $P$  i  $Q$ , możemy obliczyć rozbieżność pomiędzy nimi  $D_{KL}(P||Q)$  za pomocą dywergencji Kullbacka-Leiblera:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

W przypadku autoenkodera rzadkiego naszym celem jest zmierzenie rozbieżności pomiędzy docelowym prawdopodobieństwem  $p$  aktywacji neuronu w warstwie kodowania, a rzeczywistym prawdopodobieństwem  $q$  (które jest średnią aktywacją dla danych uczących). Dywergencję KL można wówczas zapisać jako:

$$D_{KL}(p||q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$



**Rysunek 1.16:** Funkcje straty rzadkości

Źródło: [6]

Po obliczeniu funkcji straty rzadkości dla każdego neuronu w warstwie kodowania, należy je zsumować i wynik dodać do funkcji kosztu. W celu regulowania względnej istotności funkcji straty rzadkości i funkcji straty rekonstrukcji, można tą pierwszą pomnożyć przez hiperparametr wagi rzadkości. Jeśli wartość tego hiperparametru będzie zbyt duża, to model pozostanie blisko rzadkości docelowej, ale jednocześnie nie będzie w stanie prawidłowo rekonstruować danych wejściowych. Przy zbyt małej wartości tego parametru, model będzie ignorował cel rzadkości [6].

### Autoenkodery z odszumianiem

Zamiast dodawać karę  $\Omega$  do funkcji kosztów, możemy uzyskać autoenkoder, który uczy się czegoś przydatnego, zmieniając składnik błędu rekonstrukcji w funkcji kosztów. Tradycyjne autoenkodery minimalizują jakąś funkcję

$$L(x, g(f(x)))$$

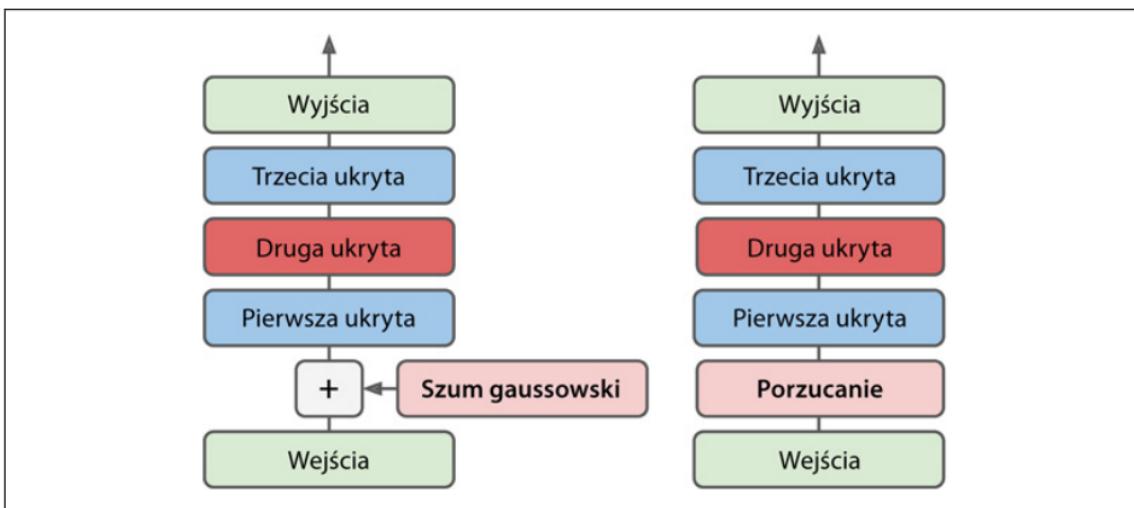
gdzie  $L$  to funkcja straty karząca  $g(f(x))$  za niepodobieństwo do  $x$ , jak np. norma  $L^2$  ich różnic. Sprzyja to temu, aby  $g \circ f$  uczyła się być jedynie funkcją tożsamościową, jeśli ma do tego odpowiednią pojemność.

Autoenkoder z odszumianiem zamiast tego minimalizuje

$$L(x, g(f(\tilde{x})))$$

gdzie  $\tilde{x}$  to kopia  $x$ , która została zniekształcona przez jakiegoś rodzaju postać szumu. Autoenkodery mają tym samym za zadanie odwrócić to zniekształcenie, a nie po prostu przekopować swoje wejście [7].

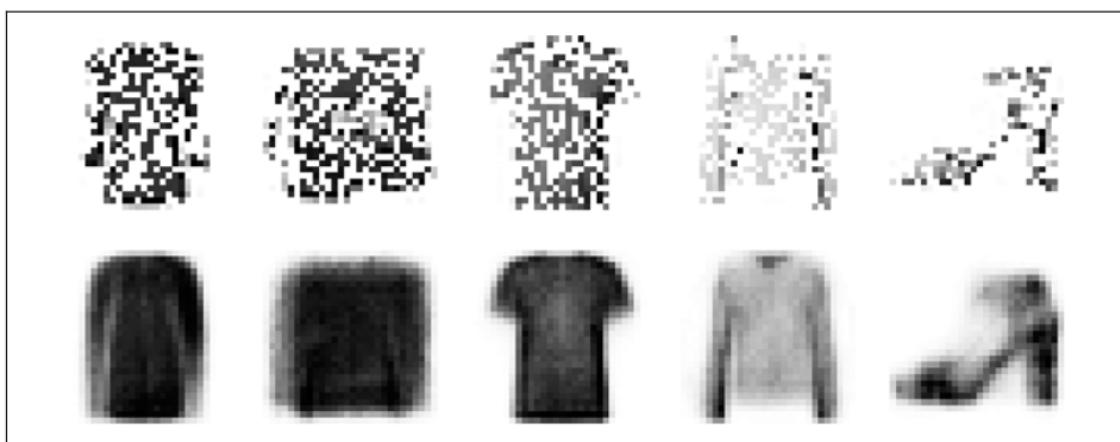
Zniekształcenie może być szumem gaussowskim dodawanym do danych wejściowych lub może przybrać postać losowo wyłączanych wejść za pomocą metody porzucania. Autoenkodery z takimi zniekształceniami zostały przedstawione na rysunku 1.17.



**Rysunek 1.17:** Autokodery odszumiające: wykorzystujące szum gaussowski (po lewej) lub metodę porzucania (po prawej)

Źródło: [6]

Rysunek 1.18 przedstawia przykłady zaszumionych obrazów (połowa pikseli została „wyłączona”), a także ich rekonstrukcje uzyskane za pomocą autokodera odszumiającego (bazującego na warstwie porzucania). Autokoder „odgaduje” szczegóły niewystępujące w obrazach wejściowych, na przykład górną część białej sukienki (czwarty obraz w dolnym rzędzie).

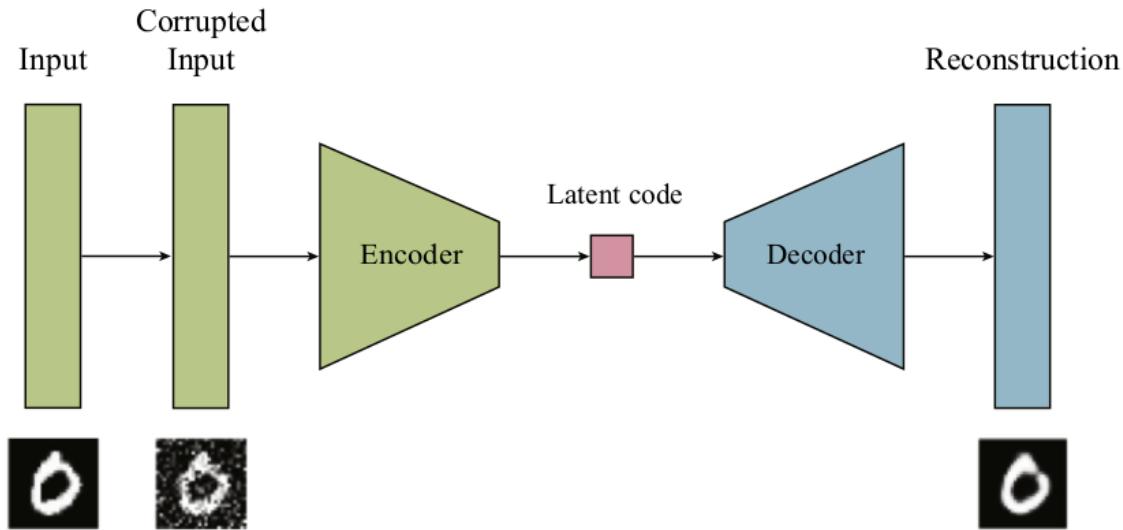


**Rysunek 1.18:** Zaszumione obrazy (na górze) i ich rekonstrukcje (na dole)

Źródło: [6]

Ogólna struktura autoenkodera odszumiającego została przedstawiona na rysunku 1.19. Wejściem jest zanieczyszczony obraz, a docelową wartością jest oryginał bez zniekształce-

nia. Sieć uczy się rozpoznawać i usuwać zniekształcenia, aby wygenerować rekonstrukcję. Autoenkoder nie widzi oryginalnego, czystego obrazu, jedynie ten ze zniekształceniem.



**Rysunek 1.19:** Struktura autoenkodera odszumiającego

### Regularizacja poprzez karanie pochodnych

Kolejną strategią regularizacji autoenkodera jest użycie kary  $\Omega$  tak, jak w rzadkich autoenkoderach

$$L(x, g(f(x))) + \Omega(h, x)$$

ale z inną postacią  $\Omega$ :

$$\Omega(h, x) = \lambda \sum_i ||\nabla_x h_i||^2$$

Zmusza to model do poznania funkcji, która nie zmienia się za bardzo przy niewielkich zmianach  $x$ . Ponieważ kara ta jest stosowana tylko w przykładach szkoleniowych, zmusza autoenkoder do poznawania cech, które przechwytyują informacje o rozkładzie szkoleniowym. Autoenkoder z taką regularizacją jest nazywany **kurczliwym** (ang. *Contractive Autoencoder*) [7]. Autoenkodery kurczliwe dzięki użyciu opisanej powyżej funkcji straty stają się odporne na niewielkie zmiany w zbiorze uczącym.

Autoenkodery kurczliwe mają pewne wspólne cechy z autoenkoderami rzadkimi oraz odszumiającymi. W autoenkoderze rzadkim celem jest, aby jak najwięcej elementów reprezentacji było bliskich zeru. W tym celu muszą one leżeć w lewej części funkcji sigmoidalnej, gdzie wartość tej funkcji jest bliska zeru, z bardzo małą pierwszą pochodną. Prowadzi to do kurczliwego odwzorowania w rzadkim autoenkoderze, mimo że nie jest to jego celem. Z kolei w przypadku autoenkodera odszumiającego, jego celem jest zwiększenie odporności kodera na małe zmiany w zbiorze uczącym, co jest podobnym celem, jak w przypadku autoenkodera

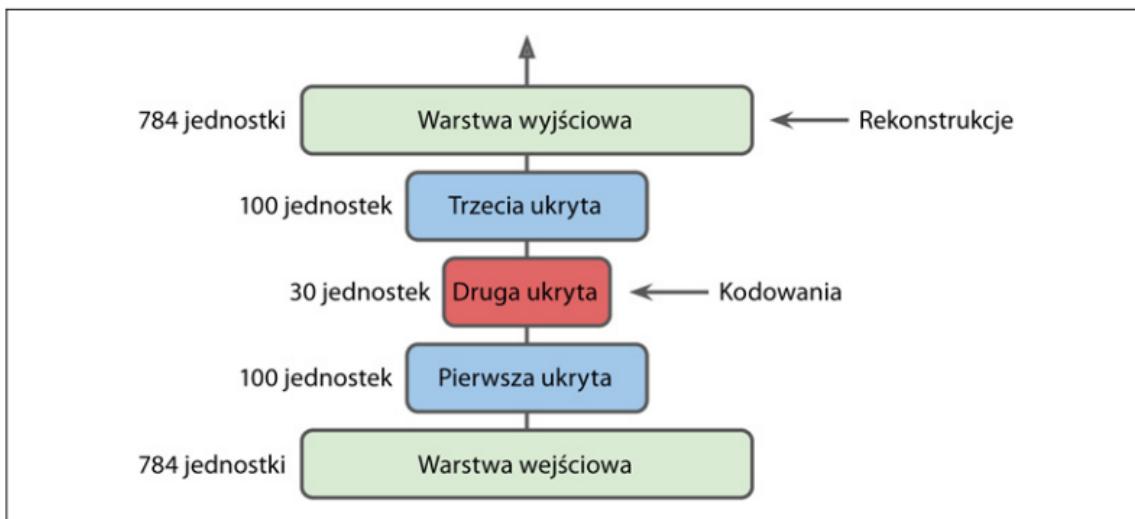
kurczliwego. Różnica polega na tym, że kurczliwe autoenkodery zwiększą odporność reprezentacji  $f(x)$ , a autoenkodery odszumiające zwiększą odporność rekonstrukcji, co tylko częściowo zwiększa odporność reprezentacji [5].

### Autoenkodery stosowe

Podobnie jak inne sieci neuronowe, również autoenkodery mogą mieć wiele warstw ukrytych. Takie autoenkodery są nazywane **stosowymi** (ang. *stacked*) lub **głębokimi** (ang. *deep*). Dzięki wielu warstwom ukrytym, autoenkoder może się uczyć bardziej skomplikowanych kodowań. Pojawia się jednak niebezpieczeństwo stworzenia modelu zbyt potężnego, podobnie jak zostało to opisane przy autoenkoderach niedopełnionych, gdzie podano teoretyczny przykład autoenkodera przekształcającego  $i$ -ty przykład uczący na pojedynczą liczbę.

Przykładowa struktura autoenkodera stosowego została przedstawiona na rysunku 1.20. Zazwyczaj architektura autoenkodera stosowego jest symetryczna względem jego centralnej warstwy ukrytej, tak jak widać na rysunku. W przypadku symetrycznych autoenkoderów, często stosowane jest **wiązanie wag** (ang. *tying weights*) warstw kodera z wagami warstw dekodera. W ten sposób liczba wag w modelu zostaje zredukowana o połowę, co przyspiesza proces uczenia i zmniejsza ryzyko przetrenowania modelu. Jeśli autoenkoder zawiera  $N$  warstw (oprócz warstwy wejściowej), a  $W_L$  oznacza wagi połączeń w  $L$ -tej warstwie, to wagi warstwy dekodera można zdefiniować jako  $W_{N-L+1} = W_L^T$  (dla  $L = 1, 2, \dots, N/2$ ).

W celu zaoszczędzenia czasu można zamiast uczyć cały autoenkoder stosowy na raz, trenować pojedyncze składowe osobno, a na koniec połączyć je w całość. To rozwiązanie nazywane jest „zachłannym uczeniem warstwowym”. Obecnie jest rzadko stosowane [6].



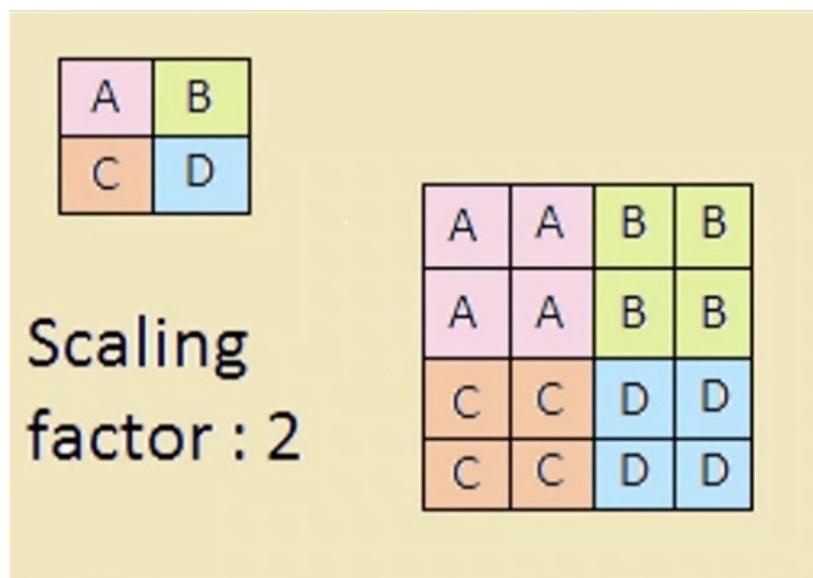
Rysunek 1.20: Przykładowa struktura autoenkodera stosowego

Źródło: [6]

### Autoenkoder splotowy

Autoenkodery splotowe są typem autoenkoderów, który najlepiej nadaje się do przetwarzania obrazów. W takim rodzaju autoenkodera, koderem jest sieć konwolucyjna składająca się z warstw splotowych i redukujących. Zwykle zmniejsza ona wymiarowość obrazu (wysokość i szerokość), a zwiększa głębokość (liczbę map cech). Dekoder przeprowadza operację odwrotną: musi zwiększyć rozdzielcość i zredukować głębokość do pierwotnego wymiaru. W tym celu można wykorzystać transponowane warstwy splotowe lub łączyć warstwy ekspansji (*UpSampling*) ze zwykłymi warstwami splotowymi. [6].

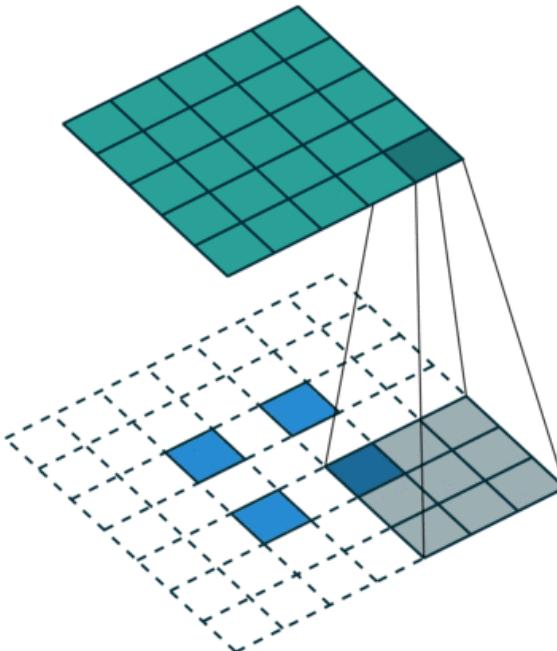
**Warstwa ekspansji** wykonuje nadpróbkowanie poprzez powtórzenie każdej wartości  $m \times n$  razy, gdzie  $m$  to współczynnik nadpróbkowania dla wierszy, a  $n$  to współczynnik nadpróbkowania dla kolumn. Przykład działania warstwy ekspansji widać na rysunku 1.21.



Rysunek 1.21: Działanie warstwy ekspansji (nadpróbkowania)

Źródło: [2]

Można powiedzieć, że **transponowane warstwy splotowe** zachowują się jak zwykłe warstwy splotowe z ułamkową długością kroku, tzn. stosują filtr do obszaru, który jest mniejszy niż rozmiar filtru. W ten sposób transponowane warstwy konwolucyjne wykonują operację w przeciwnym kierunku niż zwykłe warstwy konwolucyjne [15]. Działanie transponowanej warstwy splotowej widać na rysunku 1.22.



**Rysunek 1.22:** Działanie warstwy splotowej transponowanej

Źródło: [18]

Przy okazji autoenkoderów splotowych, warto wspomnieć o funkcji straty służącej typowo do porównania obrazów. Funkcja ta opiera się na wskaźniku podobieństwa strukturalnego (ang. *Structural Similarity Index, SSIM*). Mierzy on podobieństwo obrazów, biorąc pod uwagę trzy czynniki: jasność (ang. *luminance*), kontrast i strukturę. Niech  $\mathbf{x}$  oraz  $\mathbf{y}$  będą porównywany fragmentami obrazów o wymiarze  $N$ . Jasność jest szacowana jako średnia intensywność:

$$\mu_x = \frac{1}{N} \sum_{i=1}^N x_i$$

Funkcja porównania jasności  $l(\mathbf{x}, \mathbf{y})$  jest zatem funkcją  $\mu_x$  i  $\mu_y$ .

Następnie usuwamy średnią intensywność z porównywanych fragmentów:  $\mathbf{x} - \mu_x$ , co odpowiada rzutowi wektora  $\mathbf{x}$  na hiperprzestrzeń zdefiniowaną przez  $\sum_{i=1}^N x_i = 0$ . Używamy odchylenia standardowego jako estymatora kontrastu:

$$\sigma_x = \left( \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)^2 \right)^{\frac{1}{2}}$$

Funkcja porównania kontrastu  $c(\mathbf{x}, \mathbf{y})$  jest więc funkcją porównującą  $\sigma_x$  i  $\sigma_y$ .

W trzecim kroku wartości sygnałów są normalizowane (dzielone przez odchylenie standardowe), po to, aby dwa porównywane obrazy miały jednostkową wariancję. Porównanie struktury  $s(\mathbf{x}, \mathbf{y})$  jest przeprowadzane na tych znormalizowanych sygnałach  $(\mathbf{x} - \mu_x)/\sigma_x$  oraz  $(\mathbf{y} - \mu_y)/\sigma_y$ .

Na koniec te trzy składniki są łączone, aby uzyskać ogólny wskaźnik podobieństwa

$$S(\mathbf{x}, \mathbf{y}) = f(l(\mathbf{x}, \mathbf{y}), c(\mathbf{x}, \mathbf{y}), s(\mathbf{x}, \mathbf{y}))$$

Badane trzy składniki są stosunkowo niezależne, tzn. przykładowo zmiana jasności i/lub kontrastu nie wpłynie na zmianę struktury obrazów.

Funkcję porównania jasności definiujemy jako:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1},$$

gdzie stała  $C_1$  jest dodana, aby uniknąć niestabilności w przypadku gdy  $\mu_x^2 + \mu_y^2$  byłoby bardzo bliskie zeru. Zwykle jest ona określona jako

$$C_1 = (K_1 L)^2,$$

gdzie  $L$  jest zakresem wartości pikseli (np. 255 dla czarno-białych obrazów w zapisie 8-bitowym), a  $K_1 \ll 1$  jest pewną małą stałą.

Funkcja porównania kontrastu przybiera podobną postać:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2},$$

gdzie  $C_2 = (K_2 L)^2$ , a  $K_2 \ll 1$ .

Porównanie struktury jest przeprowadzane po odjęciu jasności i podzieleniu przez kontrast. Definiujemy je następująco:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

gdzie  $C_3$  podobnie jak  $C_1$  i  $C_2$  jest pewną niewielką stałą, a  $\sigma_{xy}$  jest estymowane następująco:

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)$$

Łącząc  $l$ ,  $c$  i  $s$  w jedną funkcję, otrzymujemy

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^\alpha \cdot [c(\mathbf{x}, \mathbf{y})]^\beta \cdot [s(\mathbf{x}, \mathbf{y})]^\gamma$$

gdzie  $\alpha, \beta, \gamma$  są dodatnimi parametrami pozwalającymi na ustalenie ważności poszczególnych składników funkcji. Zazwyczaj przyjmuje się  $\alpha = \beta = \gamma = 1$  oraz  $C_3 = C_2/2$ . Wtedy funkcja SSIM przyjmuje postać:

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

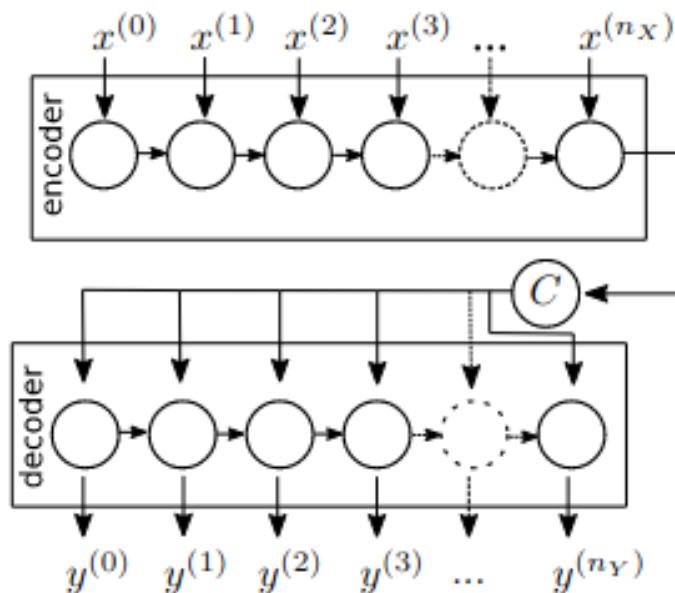
Funkcja SSIM określona powyższym wzorem, została zdefiniowana w pracy [21]. W pracy tej przyjęto  $K_1 = 0.01$  i  $K_2 = 0.03$ . Funkcja ta działa poprawnie jedynie dla obrazów w

skali szarości. Dla obrazów kolorowych, należy obliczyć wskaźnik podobieństwa strukturalnego dla każdego z kanałów barwnych, a ostateczna wartość podobieństwa jest średnią z tych wskaźników. Funkcja SSIM przyjmuje wartości mniejsze lub równe 1, przy czym wartość 1 oznacza, że  $\mathbf{x} = \mathbf{y}$ .

*Uwaga 1.6.* Funkcja SSIM sama w sobie nie jest funkcją straty - jej wynikiem jest miara podobieństwa dwóch obrazów. Jako funkcji straty można używać  $1 - \text{SSIM}(\mathbf{x}, \mathbf{y})$ .

### Autoenkoder rekurencyjny

Autoenkodery rekurencyjne zwykle służą do przetwarzania danych sekwencyjnych, takich jak szeregi czasowe czy tekst. W przypadku takiego autoenkodera, koderem zwykle jest sieć sekwencyjno-wektorowa która kompresuje sekwencję wejściową do pojedynczego wektora. Dekoderem jest sieć wektorowo-sekwencyjna odwracająca tą operację [6].



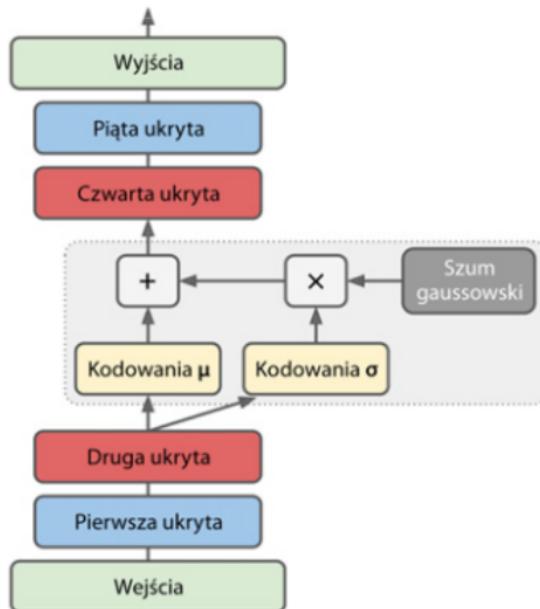
Rysunek 1.23: Przykład autoenkodera rekurencyjnego

Źródło: [20]

Przykład autoenkodera rekurencyjnego jest przedstawiony na rysunku 1.23. Zgodnie z oznaczeniami przyjętymi na tym rysunku, autoenkoder rekurencyjny generuje sekwencję wyjściową  $Y = (y^{(0)}, y^{(1)}, \dots, y^{(n_Y-1)})$  dla danej sekwencji wejściowej  $X = (x^{(0)}, x^{(1)}, \dots, x^{(n_X-1)})$ , gdzie  $n_X$  i  $n_Y$  są rozmiarami sekwencji wejściowej i wyjściowej (mogą być takie same lub różne). Zazwyczaj  $X = Y$ , aby wymusić na autoenkoderze nauczenie się semantycznego znanego danych. Na początku sekwencja wejściowa jest kodowana przez koder będący rekurencyjną siecią neuronową, a następnie reprezentacja ukryta  $C$  o danym rozmiarze jest dekodowana przez dekoder (zazwyczaj również będący rekurencyjną siecią neuronową) [20].

### Autoenkoder wariancyjny

Autoenkodery wariancyjne istotnie różnią się od opisywanych wcześniej rodzajów autoenkoderów. Są one autoenkoderami probabilistycznymi, czyli generują częściowo losowe wyniki. Stanowią klasę modeli generatywnych, co oznacza, że są w stanie tworzyć nowe dane przypominające te ze zbioru uczącego.



**Rysunek 1.24:** Przykładowa struktura autoenkodera wariancyjnego

Źródło: [6]

Na rysunku 1.24 przedstawiona jest przykładowa struktura autoenkodera wariancyjnego. Można zauważyć tutaj elementy podstawowej architektury autoenkoderów: koder i dekoder składają się z dwóch warstw ukrytych. Mamy też do czynienia z pewną modyfikacją: koder nie generuje bezpośredniego kodowania próbki wejściowej, lecz **uśrednione kodowanie**  $\mu$  oraz odchylenie standardowe  $\sigma$ . Rzeczywiste kodowanie jest następnie losowane z rozkładu normalnego o parametrach właśnie  $\mu$  i  $\sigma$ . Następnie dekoder w standardowy sposób dekoduje wylosowane kodowanie. W czasie uczenia funkcja kosztu zmusza kodowanie do stopniowego poruszania się po przestrzeni kodowania (nazywanej również przestrzenią ukrytą, z angielskiego *latent space*) w poszukiwaniu miejsca wewnątrz obszaru przypominającego chmurę punktów gaussowskich [6].

Funkcja kosztu autoenkodera wariancyjnego składa się z dwóch członów. Pierwszy jest tradycyjną funkcją straty rekonstrukcji, która zmusza autoenkoder do rekonstruowania danych wejściowych. Drugi element nazywany jest **funkcją straty ukrytej** (ang. *latent loss*). Sprawia on, że autoenkoder uzyskuje reprezentacje przypominające te uzyskiwanie z rozkładu normalnego. Stosujemy w tym celu dywergencję Kullbacka-Leiblera pomiędzy docelowej

wym rozkładem (normalnym) a rzeczywistym rozkładem kodowań. Funkcja straty ukrytej wygląda wówczas następująco:

$$-\frac{1}{2} \sum_{i=1}^n [1 + \log(\sigma_i^2) - \sigma_i^2 - \mu_i^2],$$

gdzie  $n$  - wymiarowość kodowań,  $\mu_i$  i  $\sigma_i$  to średnia i odchylenie standardowe  $i$ -tej składowej kodowania. Koder generuje na wyjściu wektory  $\mu$  i  $\sigma$ , przechowujące wszystkie wartości  $\mu_i$  i  $\sigma_i$ .

Często spotykaną modyfikacją jest zastąpienie  $\sigma$  na wyjściu kodera przez  $\gamma = \log(\sigma^2)$ . Rozwiązanie to jest stabilniejsze numerycznie i przyspiesza proces uczenia [6]. Funkcja straty ukrytej ma wówczas postać:

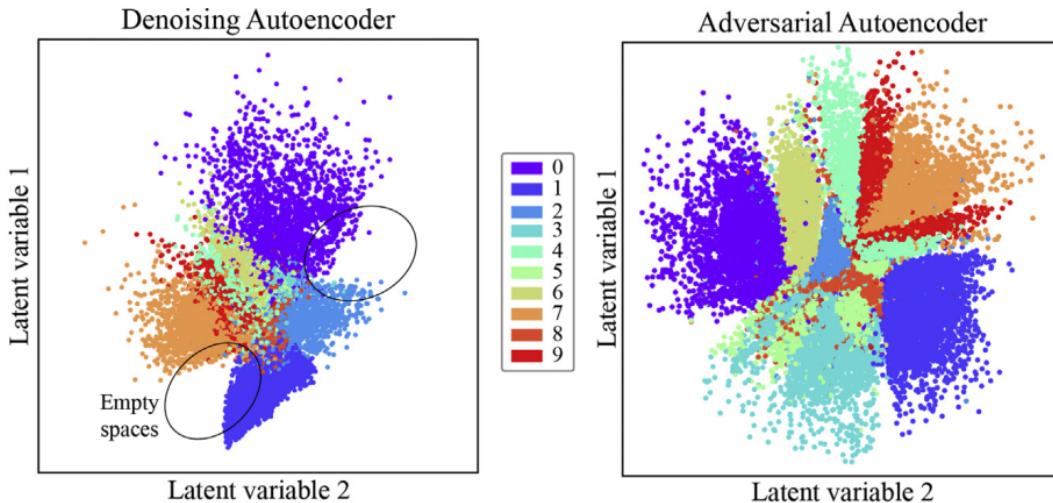
$$-\frac{1}{2} \sum_{i=1}^n [1 + \gamma_i - \exp(\gamma_i) - \mu_i^2].$$

Warto zauważyć, że po wytrenowaniu autoenkodera wariancyjnego generowanie nowych próbek jest bardzo łatwe - wystarczy wylosować kodowanie z rozkładu normalnego, a następnie rozkodować je przy użyciu dekodera [6].

### **Autoenkoder przeciwnostawny**

Ograniczeniem wielu rodzajów autoenkoderów jest to, że nie generują one dobrze ustrukturyzowanej przestrzeni ukrytej. Dzieje się tak dlatego, że podczas uczenia, różne obserwacje ze zbioru zostają zakodowane w sposób losowo rozproszony w przestrzeni ukrytej. W rezultacie rozmieszczenie reprezentacji ukrytych czasami zawiera puste miejsca w przestrzeni ukrytej, co widać na rysunku 1.25. W tych pustych miejscach znajdują się kodowania  $h$ , których dekoder nigdy nie uczył się rekonstruować. Dlatego też proces generowania nowych danych z losowo wybranego kodowania  $h$  może być trudny. W praktyce rekonstrukcja próbek z tych pustych miejsc zwykle jest błędna. Ogranicza ona zastosowanie niektórych rodzajów autoenkoderów jako modeli generatywnych [16].

Jedną z metod pozwalających na kontrolowanie struktury przestrzeni ukrytej jest stosowanie autoenkodera przeciwnostawnego. Ten rodzaj autoenkodera jest połączeniem ogólnego modelu autoenkodera z podejściem trenowania przeciwnostawnego, typowego dla generatywnych sieci przeciwnostawnych (GAN). Generatywne sieci przeciwnostawne składają się z dwóch sieci neuronowych, które konkurują ze sobą, aby polepszyć swoje działanie (rysunek 1.26). Pierwsza sieć to generator, który na wejściu dostaje losowe wartości i stara się wygenerować sztuczne dane, możliwe jak najbardziej podobne do prawdziwych danych. Drugą siecią jest dyskryminator, który decyduje, czy jego wejście pochodzi z generatora czy z prawdziwego zbioru. Celem generatora jest „oszukać” dyskryminator. Obie sieci są trenowane równocześnie, a generator nie ma bezpośredniego dostępu do prawdziwego zbioru danych. Jedynym

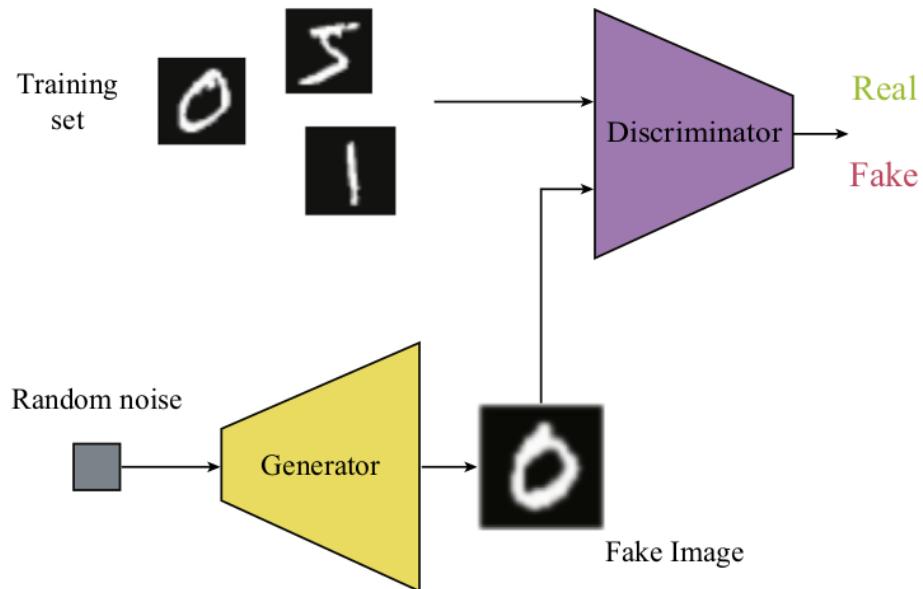


**Rysunek 1.25:** Reprezentacje ukryte. W tym przykładzie, obrazy cyfr od 0 do 9 były użyte do trenowania autoenkodera odszumiającego i przeciwnego. W obu przypadkach dane uczące są przedstawione w dwuwymiarowej przestrzeni ukrytej. Ponieważ nie nakładamy żadnych ograniczeń na autoenekoder odszumiający, to jego przestrzeń ukryta zawiera puste miejsca. Z drugiej strony, autoenekoder przeciwny ogranicza reprezentacje do podobnych według rozkładu (w tym przypadku dwuwymiarowego normalnego), co w rezultacie oznacza brak pustych miejsc

Źródło: [16]

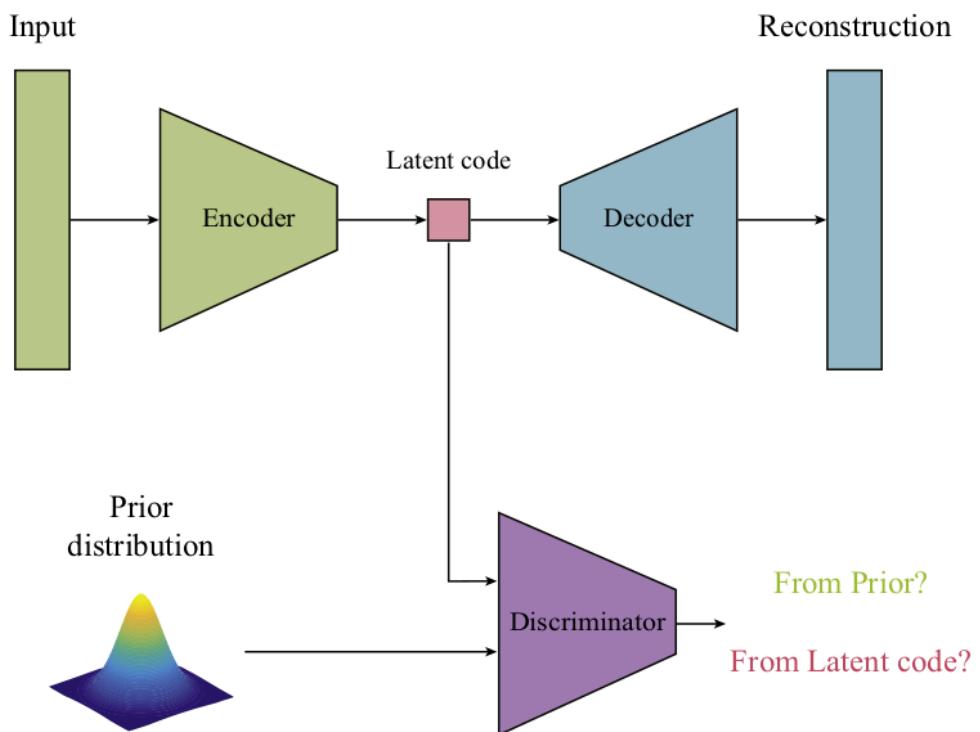
jego sposobem na poprawę działania jest interakcja z dyskryminatorem. Na podstawie informacji zwrotnych z dyskryminatora, generator stara się dopasować swoją sieć tak, by produkować wyjścia będące dla dyskryminatora trudne do zidentyfikowania jako fałszywe, czyli bardziej podobne do realnych danych.

Autoenkodery przeciwny wykorzystują trenowanie przeciwny, by ukształtować rozkład danych wejściowych tak, aby był jak najbardziej podobny do predefiniowanego rozkładu. Jest to osiągane poprzez dodanie sieci dyskryminatora do struktury autoenkodera (rysunek 1.27). W przypadku autoenkodów przeciwnych, dyskryminator otrzymuje dwa rodzaje danych wejściowych: wartości próbkiowane z pożądanego rozkładu (przykładowo losowe wartości z rozkładu normalnego) oraz ukryte reprezentacje  $h$  obserwacji ze zbioru uczącego. Co ważne, oba muszą mieć takie same wymiary. Pożądany rozkład może być tu traktowany jako rozkład a priori. Podczas procesu uczenia, dyskryminator dokonuje klasyfikacji dotyczącej tego, czy dane pochodzą z rozkładu a priori czy z kodowań ukrytych.



**Rysunek 1.26:** Struktura generatywnej sieci przeciwnostawnej (GAN)

Źródło: [16]



**Rysunek 1.27:** Struktura autokodera przeciwnostawnego. Sieć dyskryminująca jest dodana do autoenkodera, aby zmusić go do generowania reprezentacji ukrytych podobnych do rozkładu a priori

Źródło: [16]

Gdy w strukturze znajduje się dyskryminator, enkoder jest zmuszony robić jednocześnie dwie rzeczy:

(i) generować przestrzeń ukrytą, na której może pracować dekoder w celu stworzenia rekonstrukcji wejść

(ii) generować przestrzeń ukrytą, która oszuka dyskryminator tak, aby klasyfikował kodowania jako próbki z rozkładu a priori

Innymi słowy, enkoder pełni także rolę generatora.

Trenowanie autoenkodera przeciwnego ma trzy kroki w każdej epoce:

1. Autoenkoder aktualizuje parametry enkodera i dekodera, na podstawie standardowej funkcji straty rekonstrukcji
2. Model aktualizuje parametry sieci dyskryminującej, aby rozróżnić „prawdziwe” próbki (generowane z rozkładu a priori) od wygenerowanych próbek (ukrytych reprezentacji z enkodera)
3. Enkoder na podstawie informacji zwrotnej z dyskryminatora aktualizuje swoje parametry, aby poprawić generowane reprezentacje ukryte, w celu „oszukania” dyskryminatora.

Jeśli proces uczenia się powiedzie, enkoder nauczy się przekształcić rozkład danych wejściowych do rozkładu a priori, dyskryminator nigdy nie będzie pewny, czy dane wejściowe są prawdziwe czy nie, a dekoder uczy się modelu generatywnego, który odwzorowuje narzucony rozkład a priori na rozkład danych wejściowych. Ostatecznie jest możliwe generowanie nowych danych poprzez próbkowanie z rozkładu a priori, przekazanie go do dekodera i pozwolenie mu na wygenerowanie danych.

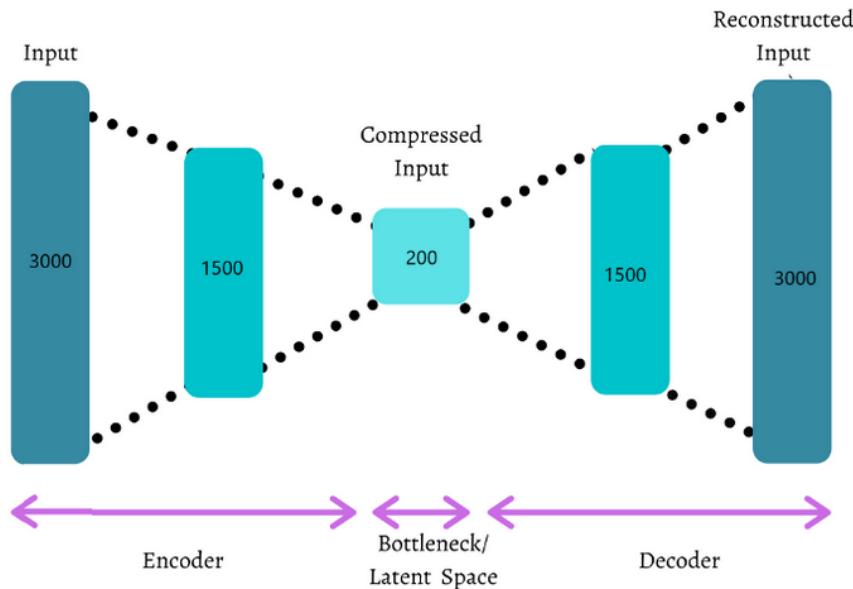
### 1.3.3. Zastosowania autoenkoderów

W zadaniach klasyfikacji lub regresji autoenkodery mogą być wykorzystywane do wyodrębniania cech z surowych danych w celu zwiększenia odporności modelu [11]. Istnieje wiele innych zastosowań sieci autoenkoderów, które można wykorzystać w różnym kontekście. Omówimy krótko zastosowania takie jak:

1. Zmniejszenie wymiarowości (ang. *Dimensionality Reduction* )
2. Wyodrębnianie cech (ang. *Feature Extraction*)
3. Odszumianie obrazu (ang. *Image Denoising* )
4. Kompresja obrazu (ang. *Image Compression*)
5. Wyszukiwanie obrazu (ang. *Image Search* )
6. Wykrywanie anomalii (ang. *Anomaly Detection*)
7. Uzupełnianie brakujących danych (ang. *Missing Value Imputation* )

### Zmniejszenie wymiarowości

Autoenkodery trenują sieć w celu wyjaśnienia naturalnej struktury danych w efektywnej reprezentacji niskowymiarowej. Osiąga się to poprzez zastosowanie strategii dekodowania i kodowania w celu zminimalizowania błędu rekonstrukcji [11].



**Rysunek 1.28:** Autodenkoder w zastosowaniu do redukcji wymiarowości

Źródło: [11]

Jak widać na rysunku 1.28, autoenkoder składa się z trzech elementów:

- koder - funkcja służąca do kompresji danych do ich reprezentacji w niższym wymiarze.
- wąskie gardło (ang. *bottleneck*) - nazywane również przestrzenią ukrytą, gdzie nasze początkowe dane są reprezentowane w niższym wymiarze.
- dekoder - funkcja dekompresji lub rekonstrukcji danych o niskim wymiarze z powrotem do wymiaru początkowego.

W przykładzie na rysunku 1.28 warstwa wejściowa i warstwa wyjściowa mają wymiar 3000, a pożądany wymiar zredukowany wynosi 200. Możemy stworzyć sieć pięciowarstwową, w której koder ma 3000 i 1500 neuronów, podobnie jak w przypadku sieci dekodera. Reprezentacje ukryte można traktować jako reprezentację o zredukowanym wymiarze.

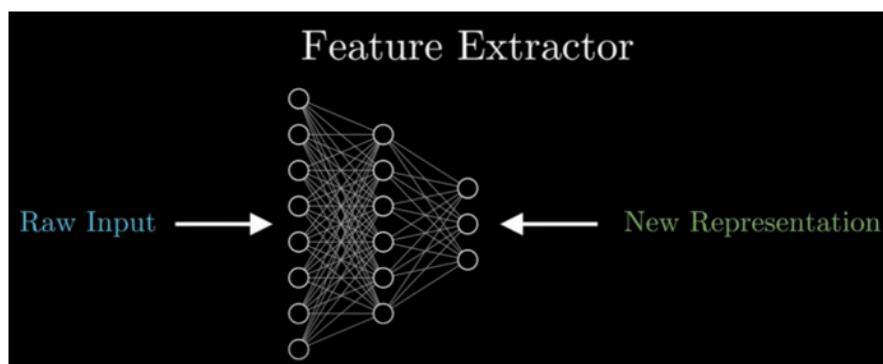
Popularną metodą redukowania wymiarowości jest analiza składowych głównych. Porównamy tą metodę ze stosowaniem autoenkoderów [14]:

- PCA jest liniową transformacją danych, podczas gdy autoenkodery mogą być liniowe lub nielinijne
- PCA jest szybsze niż autodenkodery, które są trenowane przy użyciu algorytmu spadku gradientu

- PCA gwarantuje ortogonalność wynikowej przestrzeni zmiennych, autoenkoder dąży jedynie do osiągnięcia jak najmniejszej wartości funkcji straty
- autodenkodery są w stanie modelować bardziej złożone lub nieliniowe zależności, podczas gdy PCA jest prostym przekształceniem liniowym
- przy decyzji między użyciem PCA i autodenkodera, warto wybierać PCA dla mniejszych zbiorów danych, a autodenkodery dla większych
- PCA ma tylko jeden hiperparametr - liczba ortogonalnych wymiarów, które chcemy używać. W przypadku autodenkoera występują wszystkie hiperparametry architektury sieci neuronowej
- autoenkoder z jedną warstwą i liniową funkcją aktywacji ma działanie podobne do PCA. Autoenkodery z wieloma warstwami i nieliniowymi funkcjami aktywacji (Głęboki Autoenkoder) są skłonne do przeuczenia, mogą być poprawiane przez regularyzację i odpowiednie zaprojektowanie sieci neuronowej

### Wyodrębnianie cech

Autokodery mogą być używane jako ekstraktory cech w zadaniach klasyfikacji lub regresji. Autokodery pobierają nieoznakowane dane i uczą się efektywnych kodowań struktury danych, które mogą być wykorzystane w zadaniach uczenia nadzorowanego. Po wytrenowaniu sieci autoenkodera na próbce danych treningowych można zignorować dekoder, a jedynie użyć kodera do przekształcenia surowych danych wejściowych o wyższym wymiarze na przestrzeń zakodowaną o niższym wymiarze (rysunek 1.29). Ten niższy wymiar danych może być wykorzystany jako cecha w zadaniach uczenia nadzorowanego [11].

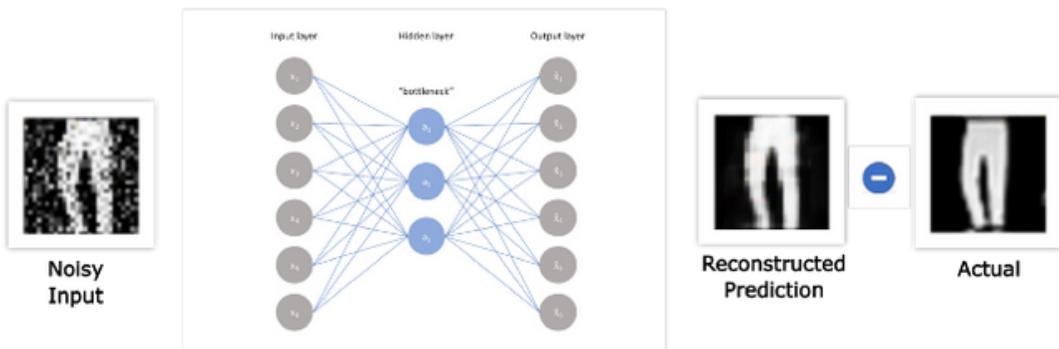


**Rysunek 1.29:** Autoenkoder stosowany do ekstrakcji cech

Źródło: [11]

## Odszumianie obrazów

Surowe dane wejściowe ze świata rzeczywistego są często zaszumione, a wytrenowanie dobrego modelu nadzorowanego wymaga danych oczyszczonych i pozbawionych szumu. Do odszumiania danych można wykorzystać autoenkodery. Jednym z popularnych zastosowań jest odszumianie obrazów, w którym autoenkodery próbują zrekonstruować obraz pozbawiony szumu z zaszumionego obrazu wejściowego [11].



**Rysunek 1.30:** Stosowanie autoenkodera do odszumiania obrazu

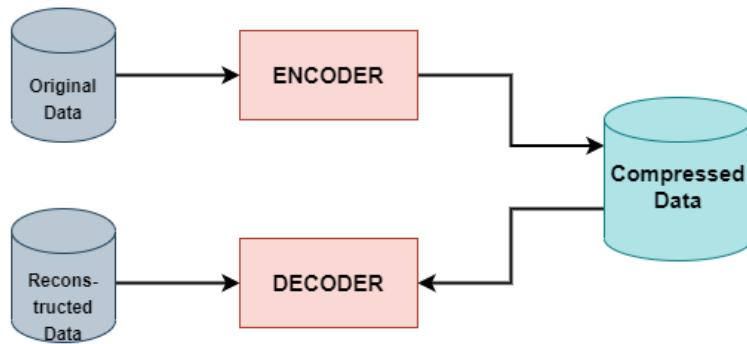
Źródło: [11]

Zakłócony obraz wejściowy jest podawany do autoenkodera jako wejście, a bezszumowe wyjście jest rekonstruowane przez minimalizację straty rekonstrukcji w stosunku do oryginalnego wyjścia docelowego (bezszumowego). Po wytrenowaniu wag autoenkodera można je dalej wykorzystać do odszumiania obrazu surowego. Schemat działania autoenkodera odszumiającego widać na rysunku 1.30.

## Kompresja obrazu

Innym zastosowaniem sieci autoenkoderów jest kompresja obrazów. Surowy obraz wejściowy można przekazać do sieci kodera i uzyskać skompresowany wymiar zakodowanych danych. Wagi sieci autoenkodera mogą być uczone przez rekonstrukcję obrazu ze skompresowanego kodowania za pomocą sieci dekodera [11]. Schemat takiego zastosowania jest pokazany na rysunku 1.31.

Zazwyczaj autoenkodery nie nadają się zbyt dobrze do kompresji danych, lepiej sprawdzają się raczej podstawowe algorytmy kompresji.

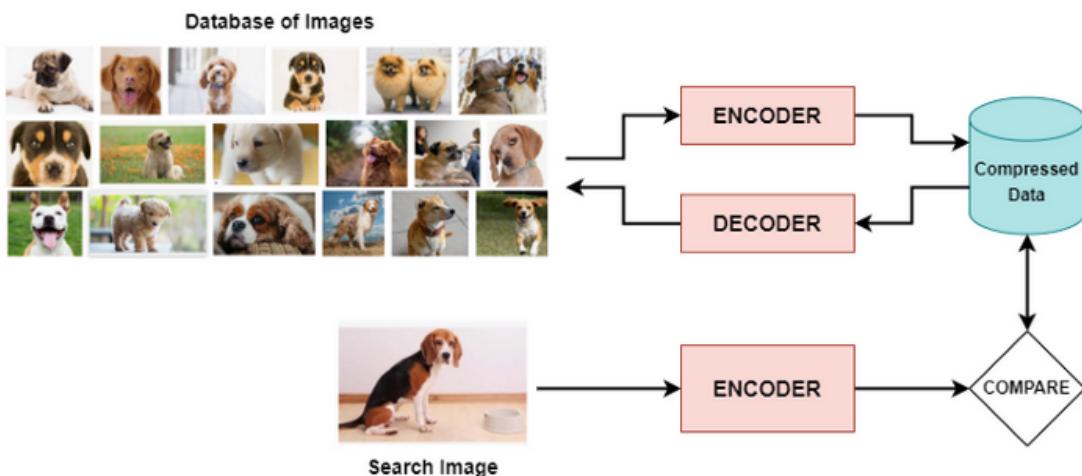


**Rysunek 1.31:** Zastosowanie autoenkodera do kompresji obrazu

Źródło: [11]

### Wyszukiwanie obrazu

Autoenkoderów można użyć do kompresji bazy danych obrazu, aby następnie skompresowana reprezentacja mogła być przy przeszukiwaniu porównywana z zakodowaną wersją szukanego obrazu [11]. W przypadku, gdy chcemy znaleźć jeden konkretny obraz w bazie danych, możemy porównywać jego kodowanie z kodowaniami pozostałych obrazów (będzie to rozwiązanie szybsze niż porównywanie całych obrazów). Takie zastosowanie jest przedstawione na rysunku 1.32. Jeżeli chcemy znaleźć  $k$  obrazów podobnych do zadanego obrazu, możemy dla ukrytych kodowań zbudować model służący do wyodrębniania podobnych elementów, jak na przykład model  $k$  najbliższych sąsiadów [22].



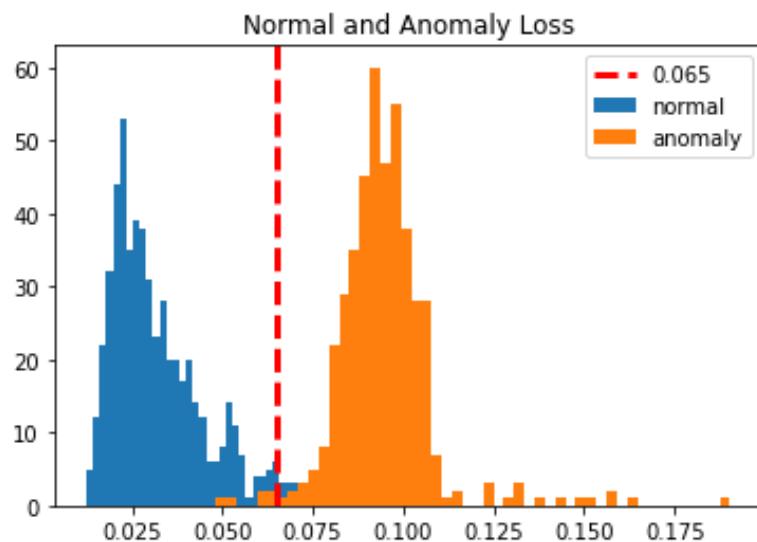
**Rysunek 1.32:** Zastosowanie autoenkodera do wyszukiwania obrazu

Źródło: [11]

### Wykrywanie anomalii

Kolejnym zastosowaniem sieci autoenkoderów jest wykrywanie anomalii. Model wykrywania anomalii można wykorzystać do wykrywania oszukańczych transakcji lub wszelkich zadań nadzorowanych o wysokim stopniu nierównowagi.

Idea polega na trenowaniu autoenkoderów tylko na próbkach danych jednej klasy (klasy większościowej). W ten sposób sieć jest w stanie zrekonstruować dane wejściowe z dobrą lub mniejszą stratą rekonstrukcji. Jeśli przez sieć autoenkodera przepuści się próbki danych innej klasy docelowej, spowoduje to porównywalnie większą stratę rekonstrukcji. Można określić wartość progową straty rekonstrukcji, której przekroczenie będzie uznawane za anomalię [11]. Wartość progowa jest zwykle określana na podstawie wartości funkcji straty dla zbioru uczącego. Przykładowo, na rysunku 1.33 widać wartości funkcji straty dla pewnego zbioru, z podziałem na obserwacje normalne (kolor niebieski) i anomalie (kolor pomarańczowy). Można zauważyć, że wartości funkcji straty są generalnie większe dla anomalii niż dla obserwacji normalnych. Jako wartość progowa dla anomalii w tym przykładzie została wybrana suma średniej i dwukrotności odchylenia standardowego z wartości funkcji straty dla zbioru uczącego [1].



**Rysunek 1.33:** Wybór wartości progowej dla anomalii w autoenkoderze na podstawie funkcji straty

Źródło: [1]

### **Uzupełnianie brakujących danych**

Do imputacji brakujących wartości w zbiorze danych można wykorzystać autoenkodery odszumiające. Idea polega na trenowaniu sieci autoenkoderów poprzez losowe umieszczenie brakujących wartości w danych wejściowych i próbę odtworzenia oryginalnych danych surowych poprzez minimalizację straty rekonstrukcji. Po wytrenowaniu wag autoenkodera rekordy zawierające brakujące wartości mogą być przepuszczane przez sieć autoenkodera w celu zrekonstruowania danych wejściowych, także z imputowanymi brakującymi cechami [11]. W imputacji braków danych można wykorzystać także autoenkodery wariancyjne, które potrafią generować dane podobne do tych ze zbioru uczącego [13].

## Rozdział 2

### Część praktyczna

Zbiorem, na przykładzie którego będziemy pokazywać zastosowania autoenkoderów, jest zbiór danych MNIST [12]. Przykładowe dane z części treningowej tego zbioru są przedstawione na rysunku 2.1. W zbiorze uczącym znajduje się 60 tysięcy obserwacji, a w testowym 10 tysięcy.

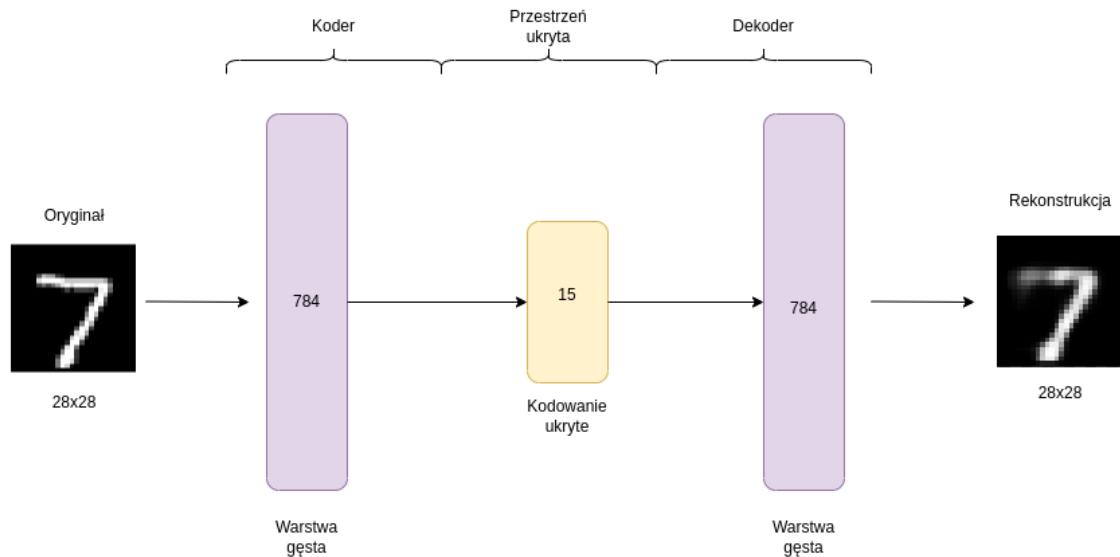


**Rysunek 2.1:** Pięć pierwszych obserwacji ze zbioru cyfr MNIST

*Źródło:* Opracowanie własne

#### 2.1. Prosty autoenkoder

Pierwszym przykładem jest tak zwany prosty autoenkoder - nazywany tak w odróżnieniu od autoenkodera splotowego. Składa się on z kodera i dekodera, z których każdy jest pojedynczą warstwą gęstą. Warstwa wyjściowa będzie przyjmować obrazy „spłaszczone” do wektora o długości 784 (obrazy 2D były wymiaru  $28 \times 28$ ). Reprezentacja ukryta ma wymiar 15, jest to więc autoenkoder niedopełniony (kodowania mają mniejszy wymiar niż dane wejściowe). Następnie dekoder przywraca wejściowy wymiar 784. Struktura użytego autoenkodera jest przedstawiona na rysunku 2.2. W tabeli 2.1 widzimy strukturę kodera, a w tabeli 2.2 strukturę dekodera.



**Rysunek 2.2:** Struktura autoenkodera prostego

Źródło: Opracowanie własne

**Tabela 2.1:** Struktura kodera w autoenkoderze prostym

Lp.	Rodzaj warstwy	Wymiar na wyjściu	Funkcja	Filtr	Krok	Uzupełnianie
1	wejściowa	784				
2	gęsta	15	ReLU			

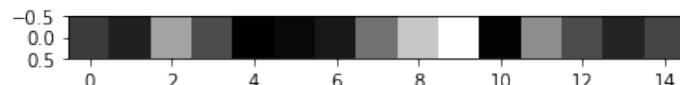
Źródło: Opracowanie własne

**Tabela 2.2:** Struktura dekodera w autoenkoderze prostym

Lp.	Rodzaj warstwy	Wymiar na wyjściu	Funkcja	Filtr	Krok	Uzupełnianie
1	wejściowa	15				
2	gęsta	784	sigmoidalna			

Źródło: Opracowanie własne

Model autoenkodera jest uczyony przez 15 epok. Rozmiar pakietu (*batch size*) wynosi 256. Zbiorem walidacyjnym jest zbiór testowy.

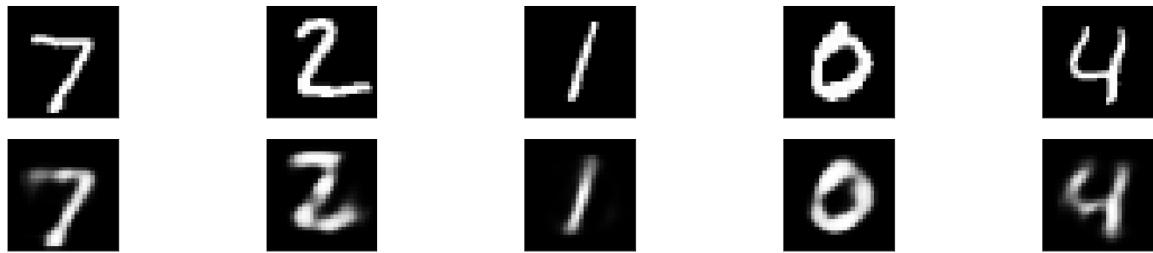


**Rysunek 2.3:** Reprezentacja 15-wymiarowa dla pierwszej obserwacji ze zbioru testowego MNIST

Źródło: Opracowanie własne

Rysunek 2.3 przedstawia kodowanie w ukrytej przestrzeni 15-wymiarowej dla pierwszej obserwacji ze zbioru testowego. Rysunek 2.4 przedstawia pięć pierwszych obrazów ze

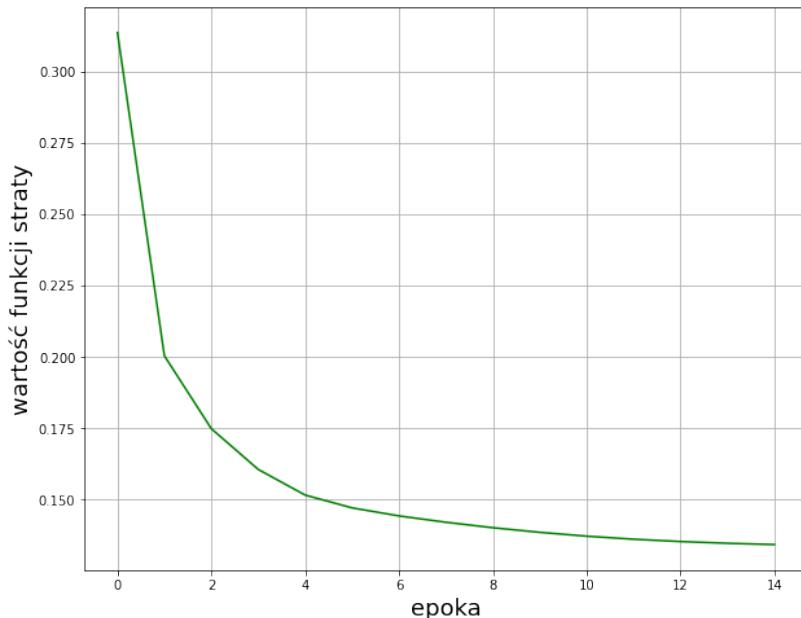
zbioru testowego (górny rząd) oraz ich rekonstrukcje po odkodowaniu przez dekoder (dolny rząd). Można zauważyć, że po przejściu przez autoenkoder niedopełniony, obrazy znacznie tracą na jakości.



**Rysunek 2.4:** Pięć pierwszych obserwacji ze zbioru testowego MNIST: w górnym rzędzie oryginalne obrazy, w dolnym rekonstrukcje z autoenkodera niedopełnionego

*Źródło:* Opracowanie własne

Rysunek 2.5 przedstawia wartość funkcji straty w kolejnych epokach uczenia autoenkodera prostego. W pierwszej epoce wartość funkcji straty wynosiła około 0.3. Widać, że największy spadek wartości funkcji straty następował w pierwszych kilku iteracjach. Po około 10 epoce wartość funkcji straty utrzymywała się na podobnym poziomie (około 0.13).



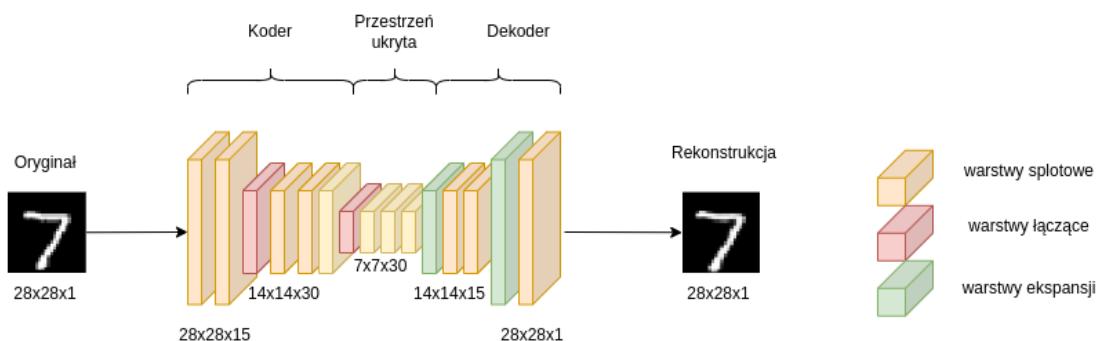
**Rysunek 2.5:** Wartość funkcji straty w kolejnych epokach trenowania autoenkodera niedopełnionego

*Źródło:* Opracowanie własne

## 2.2. Autoenkoder splotowy

W przypadku danych ze zbioru MNIST obrazy są dość małe, więc nawet prosty autoenkoder dał dość dobre wyniki. Jednak w zastosowaniu do danych wejściowych będących obrazami, zamiast autoenkodera prostego można użyć autoenkodera splotowego, który jest przeznaczony przede wszystkim do takiego typu danych. W autoenkoderze splotowym sieć kodera składa się z warstw splotowych i łączących - zazwyczaj ma ona na celu zmniejszenie wymiarowości danych (wysokości i szerokości obrazu) oraz zwiększenie głębokości (liczby map cech). Dekoder powinien przeprowadzić operację odwrotną (zwiększyć wysokość i szerokość obrazu, a zmniejszyć liczbę map cech). W tym celu można wykorzystać transponowane warstwy splotowe lub łączyć zwykłe warstwy splotowe z warstwami ekspansji [6].

Rysunek 2.6 przedstawia strukturę użytego autoenkodera splotowego. Koder składa się z dwóch warstw splotowych i dwóch warstw łączących. Warstwy łączące zmniejszają wymiary obrazu z  $28 \times 28$  kolejno do  $14 \times 14$  i  $7 \times 7$ . Dekoder składa się z dwóch warstw splotowych i dwóch warstw ekspansji, które zwiększają wymiary skompresowanego obrazu kolejno do  $14 \times 14$  i  $28 \times 28$ , osiągając oryginalny wymiar. W przypadku warstw splotowych, jądro jest wymiaru  $3 \times 3$ . Warstwy 1 i 4 mają po 15 filtrów, a warstwy 2 i 3 po 30 filtrów. Warstwa wyjściowa również jest warstwą splotową. Ma ona jeden filtr. W warstwach łączących stosowana jest metoda redukowania *MaxPooling*, gdzie jądro łączące ma rozmiar  $2 \times 2$ . W warstwach ekspansji współczynnik nadpróbkowania jest równy 2 dla wierszy i tyle samo dla kolumn. Struktura tego autoenkodera wraz z użytymi funkcjami aktywacji, rozmiarami filtrów i kroków, jest przedstawiona w tabeli 2.3 dla kodera i 2.4 dla dekodera.



**Rysunek 2.6:** Struktura autoenkodera splotowego

Źródło: Opracowanie własne

**Tabela 2.3:** Struktura kodera w autoenkoderze splotowym

Lp.	Rodzaj warstwy	Wymiar na wyjściu	Funkcja	Filtr	Krok	Uzupełnianie
1	splotowa	28x28x15	ReLU	3x3	1x1	Tak
2	łącząca	14x14x15	MaxPooling	2x2		Tak
3	splotowa	14x14x30	ReLU	3x3	1x1	Tak
4	łącząca	7x7x30	MaxPooling	2x2		Tak

Źródło: Opracowanie własne

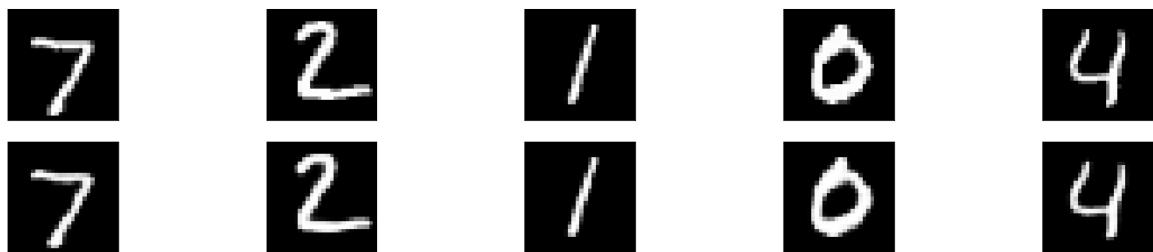
**Tabela 2.4:** Struktura dekodera w autoenkoderze splotowym

Lp.	Rodzaj warstwy	Wymiar na wyjściu	Funkcja	Filtr	Krok	Uzupełnianie
1	splotowa	7x7x30	ReLU	3x3	1x1	Tak
2	ekspansji	14x14x30	UpSampling	2x2		
3	splotowa	14x14x15	ReLU	3x3	1x1	Tak
4	ekspansji	28x28x15	UpSampling	2x2		
5	splotowa	28x28x1	sigmoidalna	3x3	1x1	Tak

Źródło: Opracowanie własne

Autoenkoder splotowy jest trenowany przez 15 epok. Rozmiar pakietu (*batch size*) wynosi 128. Zbiorem walidacyjnym jest zbiór testowy.

Na rysunku 2.7 widać efekty działania autoenkodera splotowego. W górnym rzędzie przedstawione są oryginalne obrazy ze zbioru testowego MNIST, a w dolnym rzędzie ich rekonstrukcje z autoenkodera. Widać, że obrazy w obu rzędach są do siebie podobne - strata jakości jest mniejsza niż w przypadku autoenkodera niedopełnionego.

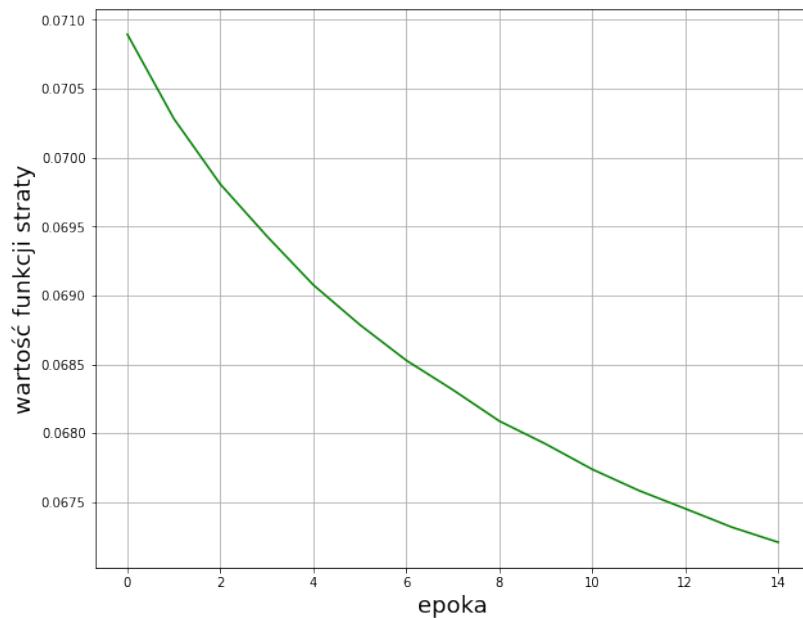


**Rysunek 2.7:** Pięć pierwszych obserwacji ze zbioru testowego MNIST: w górnym rzędzie oryginalne obrazy, w dolnym rekonstrukcje z autoenkodera splotowego

Źródło: Opracowanie własne

Rysunek 2.8 przedstawia wartości funkcji straty w kolejnych epokach trenowania autoenkodera splotowego. Widać, że wartości funkcji straty maleją przez wszystkie 15 epok, co oznacza, że raczej nie wystąpiło przeuczenie modelu. Można też zauważyć, że już w pierwszej epoce wartość funkcji straty jest niższa niż w 15 epokach w przypadku autoenkodera

niedopełnionego. Oznacza to, że rekonstrukcje z autoenkodera splotowego znacznie mniej różnią się od oryginalnych danych niż rekonstrukcje z autoenkodera niedopełnionego.



**Rysunek 2.8:** Wartość funkcji straty w kolejnych epokach trenowania autoenkodera splotowego

Źródło: Opracowanie własne

### 2.3. Autoenkoder odszumiający

Pokażemy teraz zastosowanie autoenkodera splotowego do odszumiania zanieczyszczonych obrazów. W poprzednich przykładach zbiór uczący stanowił zarówno dane wejściowe, jak i oczekiwane dane wyjściowe w modelu. W przypadku autoenkodera odszumiającego zbiór uczący również jest traktowany jako docelowy wynik, ale danymi wejściowymi są zaszumione obserwacje z tego zbioru.

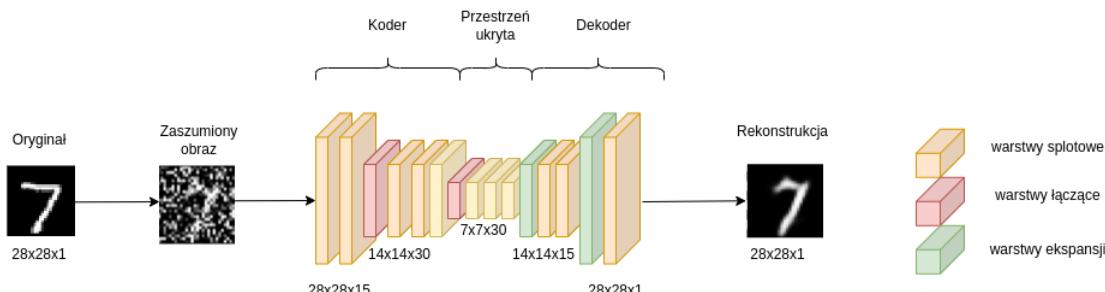


**Rysunek 2.9:** Zaszumione obrazy ze zbioru testowego MNIST

Źródło: Opracowanie własne

Rysunek 2.9 przedstawia obrazy ze zbioru testowego MNIST po zaszumieniu. Zaszumienie polegało na dodaniu do wartości każdego piksela losowej liczby z rozkładu  $\mathcal{N}(0, 1)$ , pomnożonej przez 0.7. Parametr 0.7 został przyjęty arbitralnie i można go zmieniać w celu dalszych eksperymentów, jednakże przy większej jego wartości rekonstrukcje były już znacznie mniej czytelne przy użytym autoenkoderze.

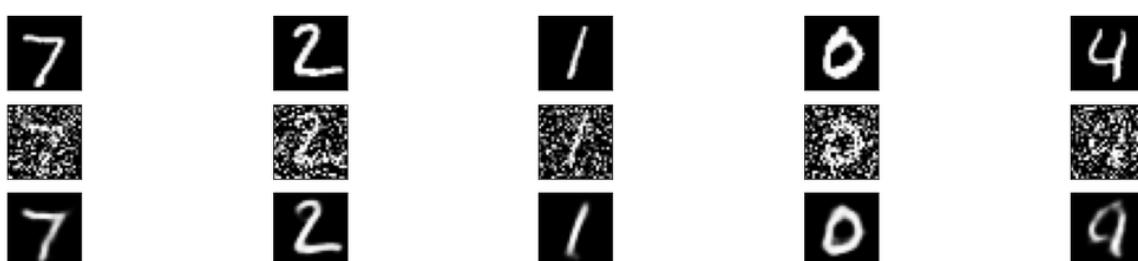
Struktura autoenkodera odszumiającego jest taka sama jak autoenkodera splotowego (rysunek 2.6 oraz tabele 2.3 i 2.4). Różnica polega jedynie na tym, że w tym przykładzie danymi wejściowymi są zaszumione obrazy, co zostało pokazane na rysunku 2.10.



**Rysunek 2.10:** Struktura autoenkodera odszumiającego

Źródło: Opracowanie własne

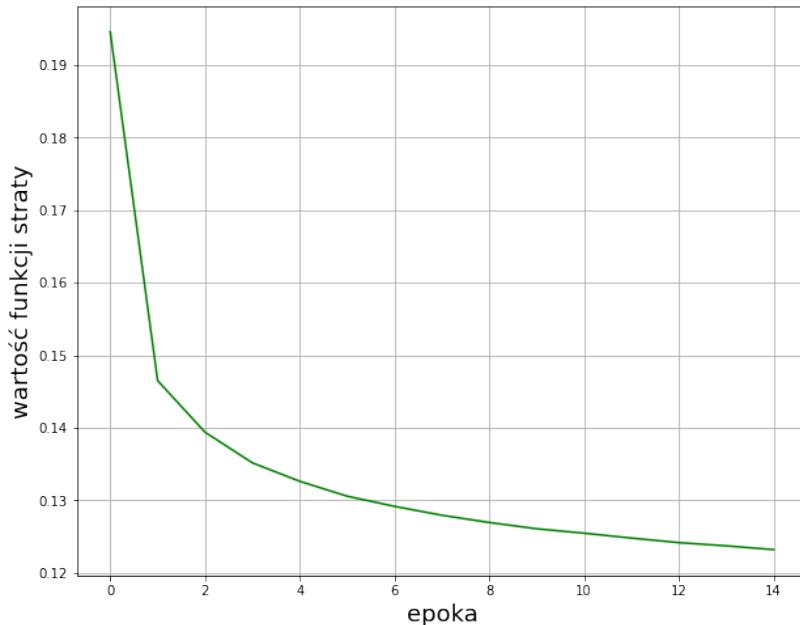
Rysunek 2.11 przedstawia efekt działania autoenkodera odszumiającego. W górnym rzędzie widać oryginalne, niezaszumione obrazy ze zbioru testowego. Środkowy rzad przedstawia zanieczyszczone obrazy, które stanowią dane wejściowe dla autoenkodera. W dolnym rzędzie widać obrazy bez szumu, zrekonstruowane przez autoenkodera. Widać, że w większości autoenkoder dobrze radzi sobie z usuwaniem zanieczyszczeń - jedynie na piątym obrazie widać w jego rekonstrukcji wyraźną pozostałość po szumie.



**Rysunek 2.11:** Pięć pierwszych obserwacji ze zbioru testowego MNIST: w górnym rzędzie oryginalne obrazy, w środkowym zaszumione, w dolnym rekonstrukcje z autoenkodera odszumiającego

Źródło: Opracowanie własne

Na rysunku 2.12 przedstawiona jest wartość funkcji straty w kolejnych epokach uczenia autoenkodera odszumiającego. Można zauważyć, że wartości funkcji straty są znacznie większe niż w poprzednim przykładzie, gdzie danymi wejściowymi były oryginalne obrazy.



**Rysunek 2.12:** Wartość funkcji straty w kolejnych epokach trenowania autoenkodera odszumiającego  
 Źródło: Opracowanie własne

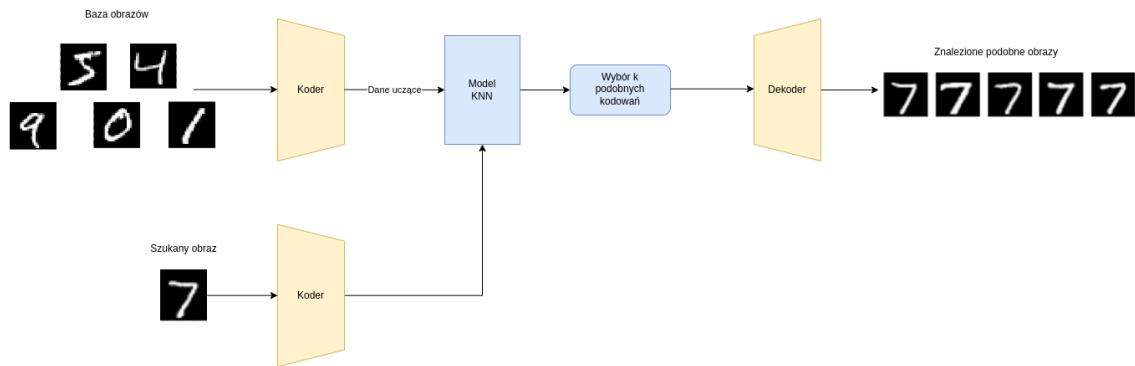
## 2.4. Wyszukiwanie obrazu

Do wyszukiwania obrazu zostanie użyty autoenkoder prosty opisany w podrozdziale 2.1. Elementy ze zbioru treningowego zostały przy użyciu kodera skompresowane do wymiaru 15. Dla skompresowanych danych budowany jest model  $k$  najbliższych sąsiadów, gdzie parametr  $k$  oznacza również liczbę podobnych obrazów do zadanego, jakie zostaną wyszukane. Zbiorem testowym są w tym przypadku obrazy, do których chcemy wyszukać obrazy podobne. Użyty zbiór testowy jest przedstawiony na rysunku 2.13. Dla każdego z tych obrazów chcemy wyszukać  $k = 5$  obrazów podobnych.



**Rysunek 2.13:** Zbiór testowy dla zadania wyszukiwania obrazu  
 Źródło: Opracowanie własne

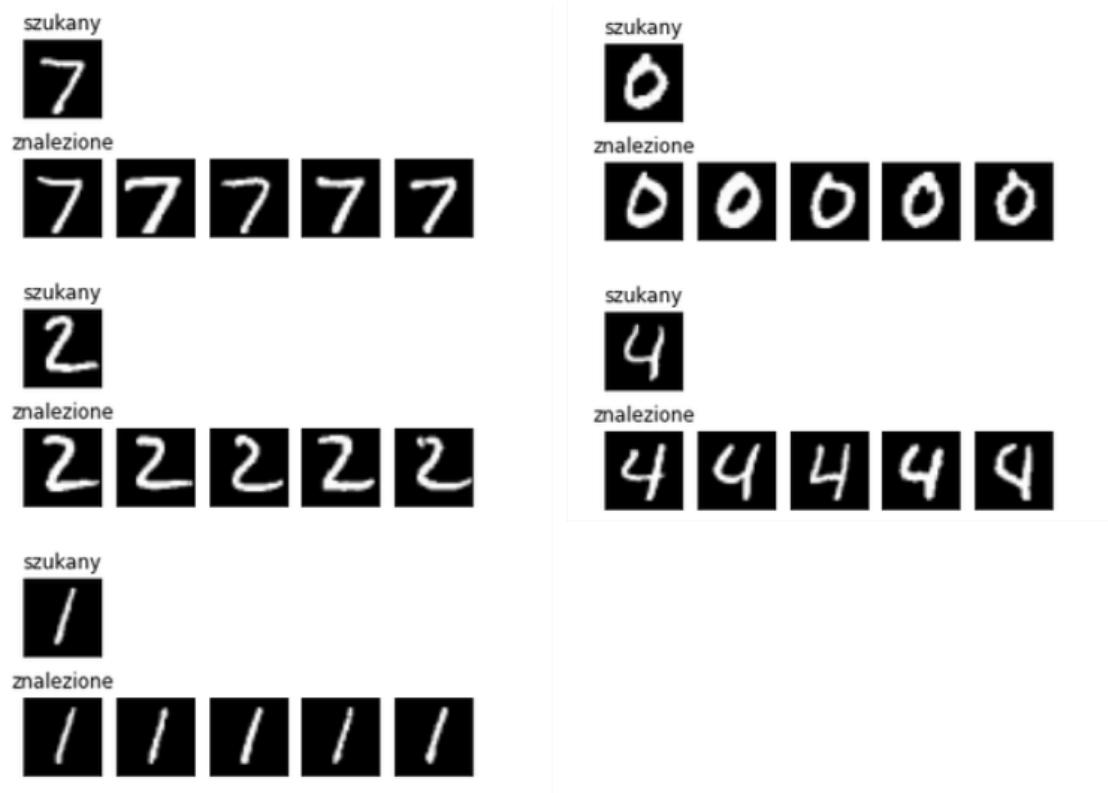
Obrazy ze zbioru testowego również są kompresowane przy użyciu kodera. Następnie dla skompresowanej formy obrazów znajdujemy  $k = 5$  najbliższych sąsiadów według modelu KNN. Jako wynik wyświetlane są oryginalne wersje znalezionych obrazów. Schemat tego zadania jest przedstawiony na rysunku 2.14.



Rysunek 2.14: Schemat wyszukiwania podobnych obrazów za pomocą autoenkodera

Źródło: Opracowanie własne

Efekt zastosowania autoenkodera oraz modelu  $k$  najbliższych sąsiadów dla opisanego zbioru testowego znajduje się na rysunku 2.15. Widać, że dla każdego elementu ze zbioru testowego, znalezione obrazy zawierają odpowiednią cyfrę, zgodną z wyszukiwaniem (np. dla pierwszego przykładu z liczbą 7, wszystkie znalezione obrazy również przedstawiają tą cyfrę).

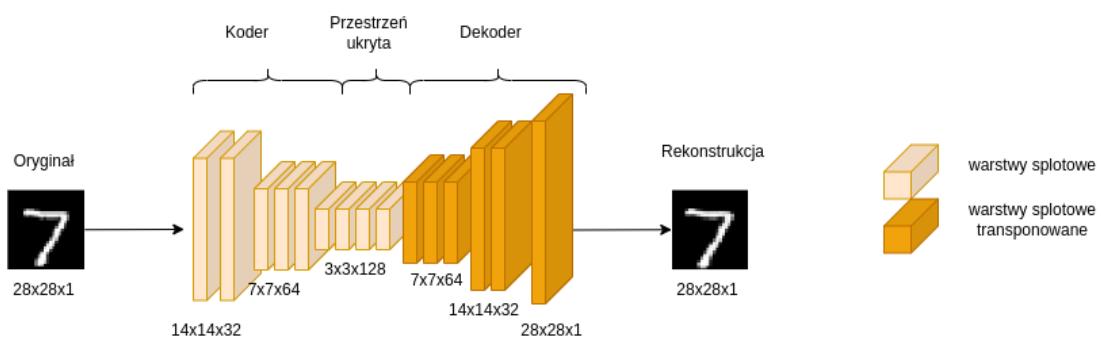


Rysunek 2.15: Wyniki zastosowania prostego autoenkodera przy wyszukiwaniu obrazu

Źródło: Opracowanie własne

## 2.5. Wykrywanie anomalii przy użyciu autoenkodera

Wykrywanie anomalii przy użyciu autoenkodera zwykle polega na tym, że sieć autoenkodera jest trenowana na zbiorze „prawidłowych” danych - bez anomalii. Następnie, jeśli w zbiorze testowym model trafi na anomalię, funkcja straty dla niej powinna mieć znacznie większe wartości niż dla obserwacji podobnych do tych „prawidłowych”. W przypadku nienadzorowanym, kiedy nie wiemy, które obserwacje są anormalne, w zbiorze uczącym mogą się znaleźć także anomalie. Na podstawie wartości funkcji straty dla obserwacji zbioru uczącego, ustalamy wartość graniczną, powyżej której przykłady testowe będą uznawane za anomalię.



**Rysunek 2.16:** Struktura autoenkodera splotowego użytego do wykrywania anomalii

Źródło: Opracowanie własne

Struktura autoenkodera splotowego wykorzystanego w zadaniu wykrywania anomalii w zbiorze MNIST jest przedstawiona na rysunku 2.16. W każdej warstwie jądro ma rozmiar 3x3, a krok wynosi 2. Funkcją aktywacji w każdej warstwie jest funkcja ReLU. Zastosowano uzupełnianie zerami. Struktura kodera jest przedstawiona w tabeli 2.5, a dekodera w tabeli 2.6. Schemat zadania wykrywania anomalii za pomocą autoenkodera jest przedstawiony na rysunku 2.17.

**Tabela 2.5:** Struktura kodera w autoenkoderze splotowym użytym do wykrywania anomalii

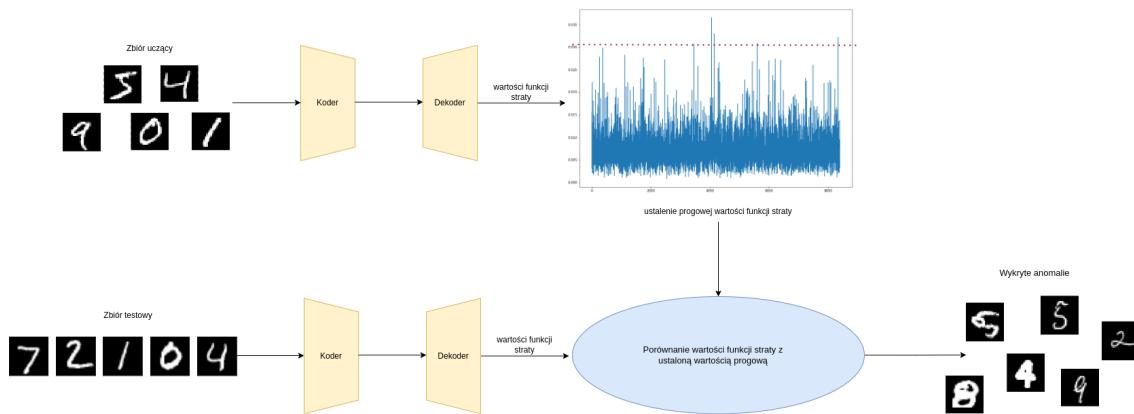
Lp.	Rodzaj warstwy	Wymiar na wyjściu	Funkcja	Filtr	Krok	Uzupełnianie
1	splotowa	14x14x32	ReLU	3x3	2x2	Tak
2	splotowa	7x7x64	ReLU	3x3	2x2	Tak
3	splotowa	3x3x128	ReLU	3x3	2x2	Nie

Źródło: Opracowanie własne

**Tabela 2.6:** Struktura dekodera w autoenkoderze splotowym użytym do wykrywania anomalii

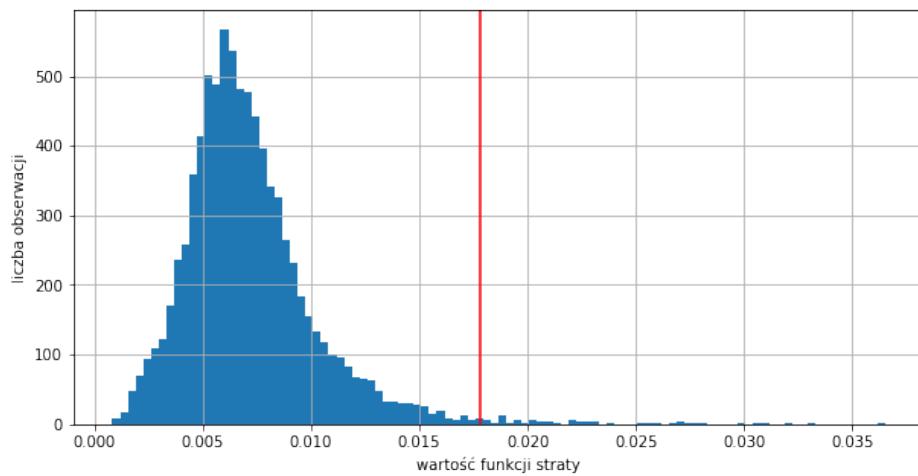
Lp.	Rodzaj warstwy	Wymiar na wyjściu	Funkcja	Filtr	Krok	Uzupełnianie
1	splotowa transponowana	7x7x64	ReLU	3x3	2x2	Nie
2	splotowa transponowana	14x14x32	ReLU	3x3	2x2	Tak
3	splotowa transponowana	28x28x1	ReLU	3x3	2x2	Tak

Źródło: Opracowanie własne


**Rysunek 2.17:** Schemat wykrywania anomalii za pomocą autoenkodera

Źródło: Opracowanie własne

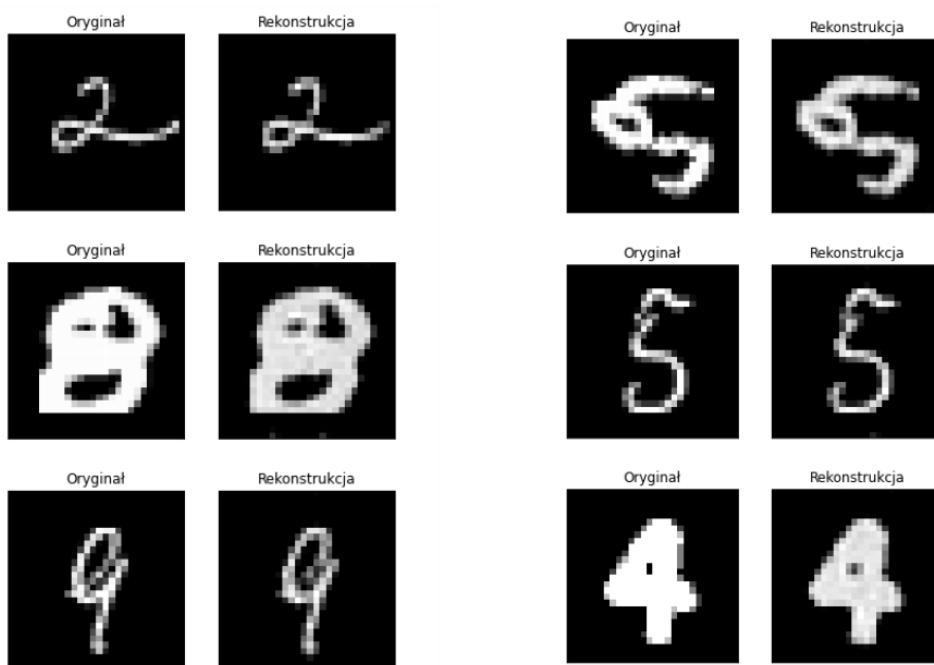
Funkcja straty bazuje na wskaźniku podobieństwa strukturalnego SSIM, a dokładniej jest to jeden minus wartość tego wskaźnika. Autoenkoder jest trenowany przez 5 epok. Na rysunku 2.18 widoczny jest histogram wartości funkcji straty dla obserwacji ze zbioru uczącego. Graniczną wartością straty, przy której uznamy obserwację za anormalną, jest dwieście dziesiąty dziewiąty percentyl wartości funkcji straty ze zbioru treningowego (wynosi on około 0.017821). Wartość ta na wykresie jest zaznaczona za pomocą pionowej czerwonej linii.



**Rysunek 2.18:** Wartości funkcji straty z autoenkodera do wykrywania anomalii, z zaznaczoną wartością progową, powyżej której obserwacja jest uznawana za anomalię

*Źródło:* Opracowanie własne

Przykłady obserwacji uznanych przez opisany model za anomalie są zaprezentowane na rysunku 2.19. Dla każdej obserwacji przedstawiono jej oryginał oraz rekonstrukcję z autoenkodera.

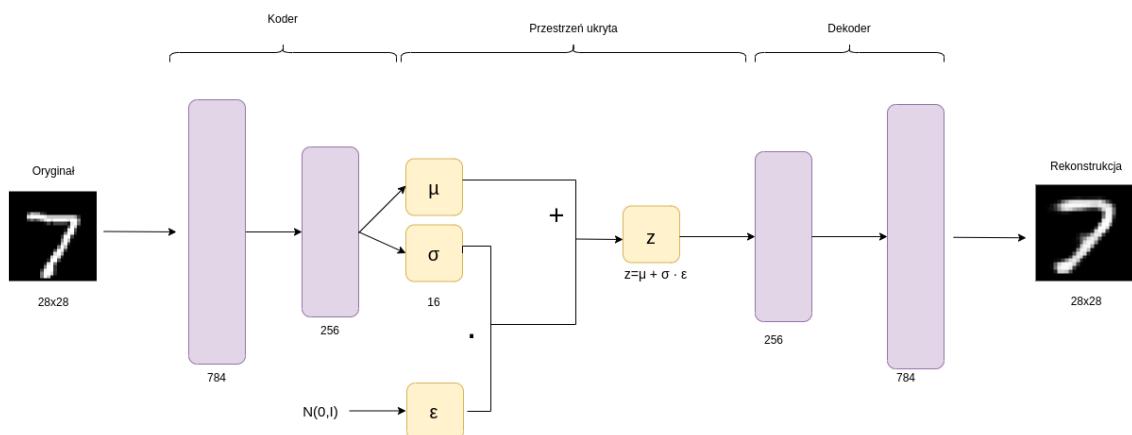


**Rysunek 2.19:** Przykłady anomalii w zbiorze MNIST wykrytych za pomocą autoenkodera

*Źródło:* Opracowanie własne

## 2.6. Generowanie obrazów przy użyciu autoenkodera wariacyjnego

Do generowania obrazów podobnych do obserwacji ze zbioru uczącego MNIST zostanie użyty autoenkoder o warstwach gęstych. Na wejściu model dostaje zatem obrazy „spłaszczone” do wektorów o długości 784. Koder zawiera jedną warstwę ukrytą, której wymiar wyjściowy to 256, a funkcją aktywacji jest ReLU. W kolejnej warstwie zachodzi kodowanie ukryte do wymiaru 16. Na podstawie wartości  $\mu$  i  $\sigma$  dokonywane jest próbkowanie z rozkładu normalnego w przestrzeni ukrytej. Wynik tego próbkowania jest przekazywany do dekodera. Dekoder składa się z dwóch warstw ukrytych (gęstych). Pierwsza zwiększa wymiar do 256, jej funkcją aktywacji jest ReLU. Druga zwiększa wymiar do oryginalnego 784, a jej funkcja aktywacji to funkcja sigmoidalna. Struktura tego autoenkodera jest przedstawiona na rysunku 2.20. Struktura kodera jest widoczna w tabeli 2.7, a dekodera w tabeli 2.8. Funkcją straty jest binarna entropia krzyżowa. Funkcją straty ukrytej jest dywergencja Kullbacka-Leiblера. Liczba epok uczenia wynosi 50, a rozmiar jednego pakietu to 128.



**Rysunek 2.20:** Struktura autoenkodera wariacyjnego

Źródło: Opracowanie własne

**Tabela 2.7:** Struktura kodera w autoenkoderze wariacyjnym

Lp.	Rodzaj warstwy	Wymiar na wyjściu	Funkcja	Filtr	Krok	Uzupełnianie
1	wejściowa	784				
2	gęsta	256	ReLU			
3	gęsta ( $\mu$ )	16				
4	gęsta ( $\sigma$ )	16				
5	gęsta ( $z$ )	16	$z = \mu + \sigma \cdot \epsilon$			

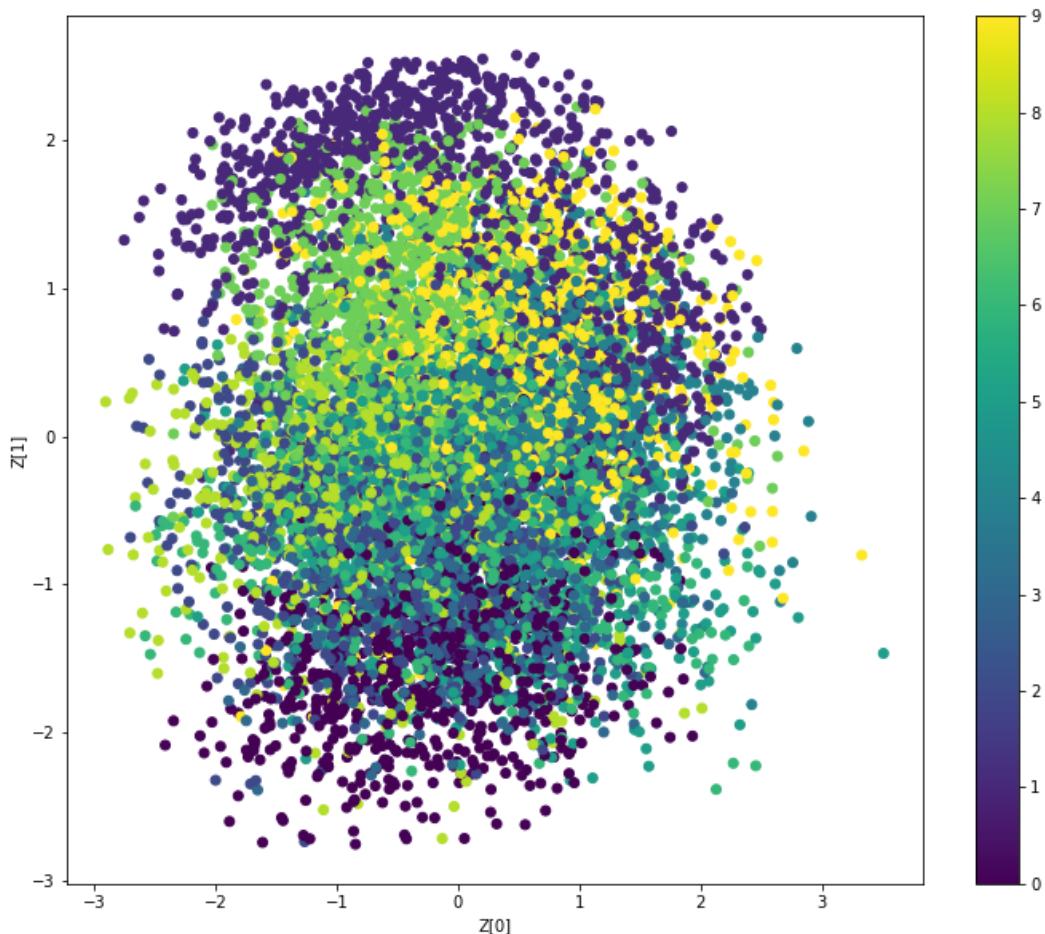
Źródło: Opracowanie własne

**Tabela 2.8:** Struktura dekodera w autoenkoderze wariancyjnym

Lp.	Rodzaj warstwy	Wymiar na wyjściu	Funkcja	Filtr	Krok	Uzupełnianie
1	wejściowa	16	ReLU	3x3	2x2	
2	gęsta	256	ReLU			
3	gęsta	784				

Źródło: Opracowanie własne

Rysunek 1.25 przedstawia pierwsze dwa wymiary przestrzeni ukrytej autoenkodera wariancyjnego użytego do generowania obrazów cyfr na podstawie zbioru MNIST. W dwóch wymiarach nie widać, aby różne kategorie obrazów były wyraźnie od siebie oddzielone.

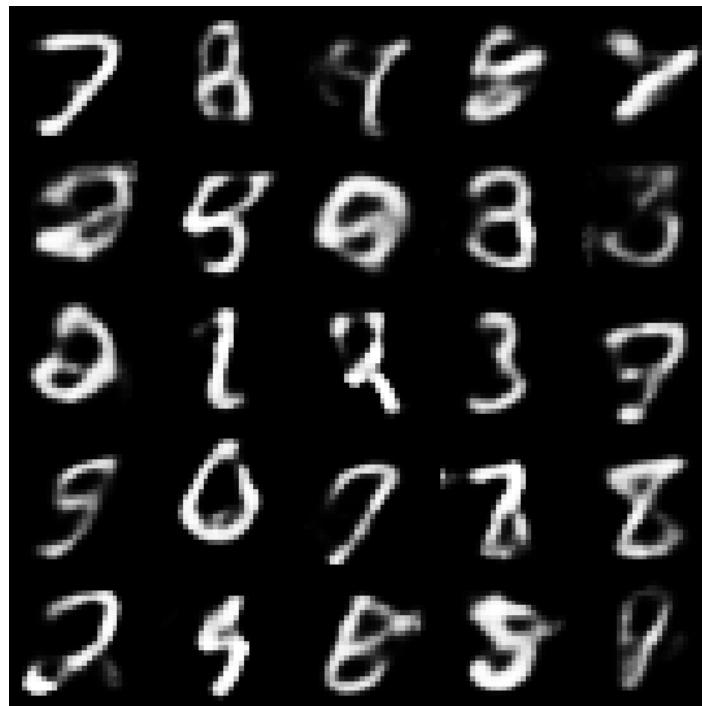


**Rysunek 2.21:** Dwa pierwsze wymiary przestrzeni ukrytej autoenkodera wariancyjnego

Źródło: Opracowanie własne

Na rysunku 2.22 widoczne są obrazy liczb wygenerowane przez autoenkoder wariancyjny. Generowanie każdego obrazu polega na wylosowaniu z rozkładu normalnego wektora o rozmiarze kodowania ukrytego (w naszym przykładzie jest to 16), a następnie na dokonaniu dla tego wektora predykcji z dekodera. W ten sposób otrzymujemy wektor długości 784, który

ry następnie przekształcany jest do wymiaru 28x28 i prezentowany w formie czarno-białego obrazu.

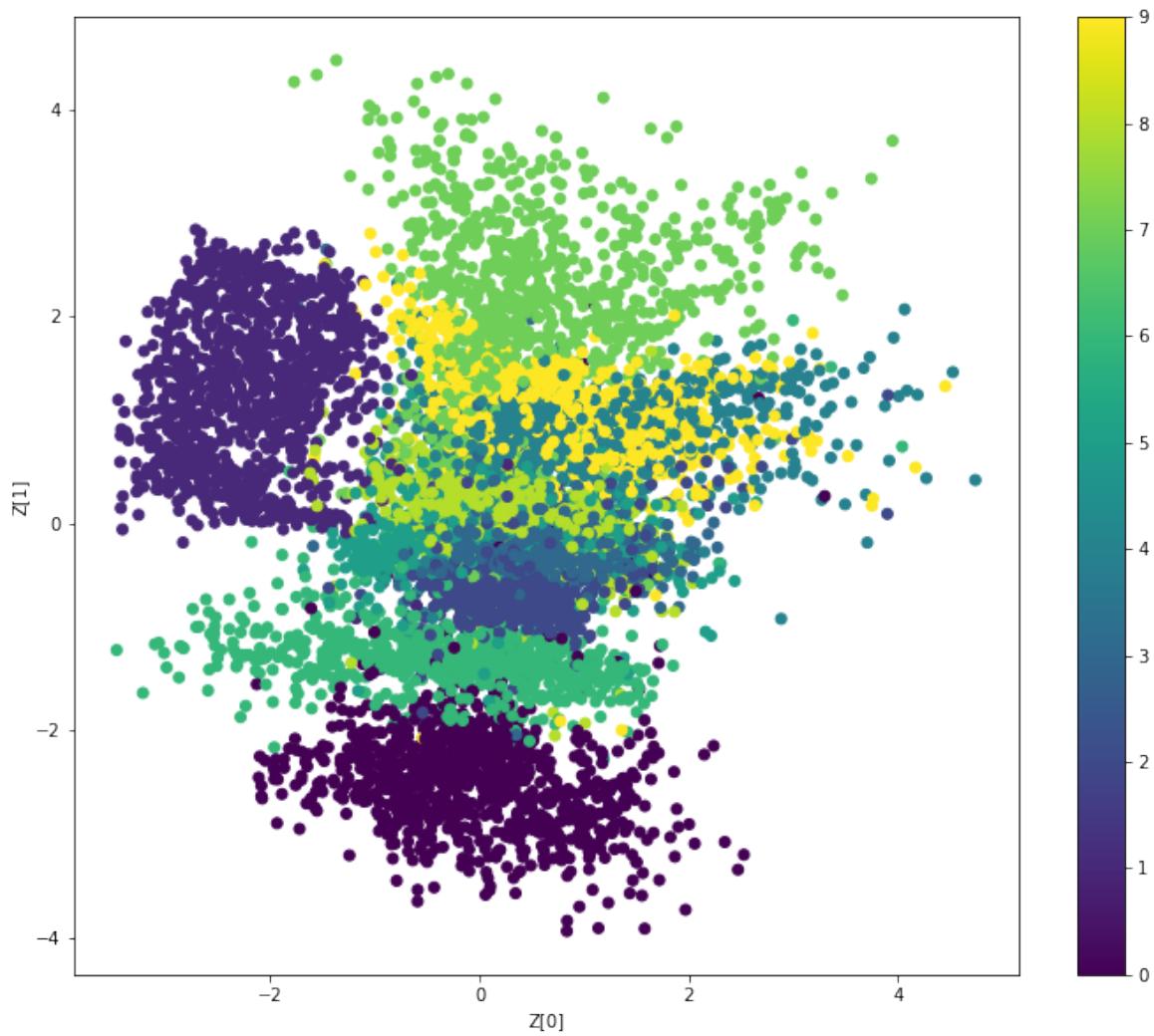


**Rysunek 2.22:** Przykłady obrazów cyfr wygenerowanych przez autoenkoder wariacyjny

Źródło: Opracowanie własne

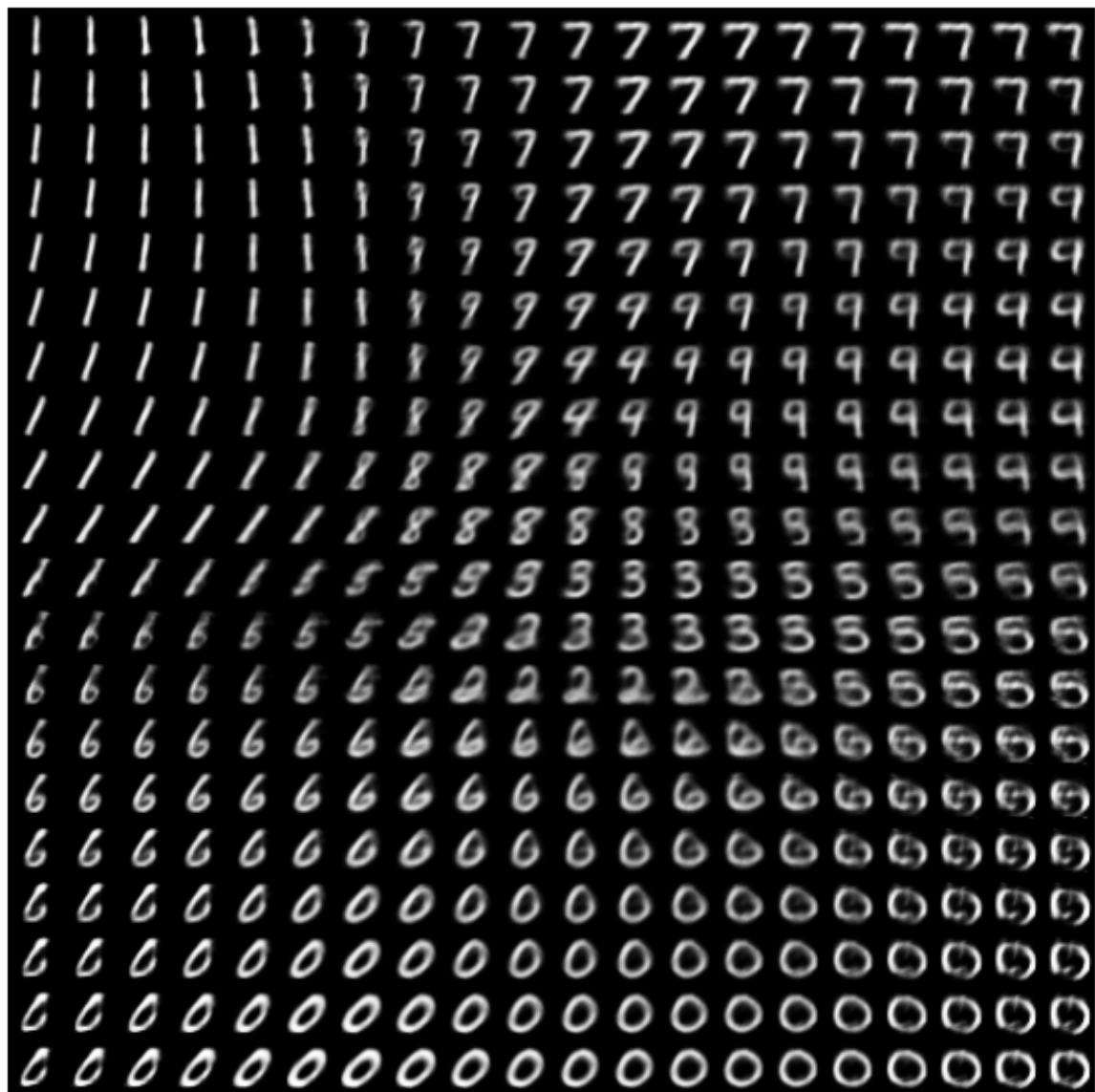
Zadanie generowania nowych obrazów zostało również przeprowadzone dla autoenkodera identycznego jak na rysunku 2.20, ale z przestrzenią ukrytą o wymiarze równym 2. Pozostałe parametry pozostają bez zmian. Przestrzeń ukryta takiego autoenkodera jest zaprezentowana na rysunku 2.23. W przypadku niektórych kategorii widać, że są wyraźnie oddzielone od pozostałych. Można również zauważać, że obserwacje nie są równomiernie rozłożone po całej przestrzeni ukrytej - problem ten został poruszony w części teoretycznej przy okazji autoenkoderów przeciwnstwowych.

Rysunek 2.24 również przedstawia przestrzeń ukrytą, ale w postaci obrazów cyfr. Możemy zaobserwować, jak jeden rodzaj cyfry „przechodzi” w inny rodzaj, na przykład w pierwszym rzędzie widzimy transzycję od liczby 1 do liczby 7.



**Rysunek 2.23:** Przestrzeń ukryta autoenkodera wariancyjnego z wymiarem kodowania ukrytego równym dwa

Źródło: Opracowanie własne



Rysunek 2.24: Przestrzeń ukryta w postaci obrazów cyfr

Źródło: Opracowanie własne



## **Podsumowanie i wnioski**

tu będzie jakieś podsumowanie a może nawet jakieś wnioski



## Bibliografia

- [1] Agrawal R., (2022), *Complete Guide to Anomaly Detection with AutoEncoders using Tensorflow* <https://www.analyticsvidhya.com/blog/2022/01/complete-guide-to-anomaly-detection-with-autoencoders-using-tensorflow/> (dostęp: 15.05.2022)
- [2] Artificial Intelligence in Plain English, *Convolutional Autoencoders (CAE) with Tensorflow* <https://ai.plainenglish.io/convolutional-autoencoders-cae-with-tensorflow-97e8d8859cbe> (dostęp: 15.05.2022)
- [3] Bank D., Koenigstein N., Giryes R. (2021), *Autoencoders*, arXiv:2003.05991v2
- [4] Ertel W. (2017) *Introduction to Artificial Intelligence. Second Edition*, Springer International Publishing
- [5] Geeks For Geeks, *Contractive Autoencoder (CAE)* <https://www.geeksforgeeks.org/contractive-autoencoder-cae/> (dostęp: 15.05.2022)
- [6] Géron A. (2020) *Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow. Wydanie II*, Helion SA
- [7] Goodfellow I., Bengio Y., Courville A. (2018), *Deep Learning. Systemy uczące się*, PWN, Warszawa
- [8] Hubel D. H., *Single Unit Activity in Striate Cortex of Unrestrained Cats*, J. Physiol. (1959) 147, 226-238
- [9] Hubel D. H., Wiesel T. N., *Receptive Fields of Single Neurones in the Cat's Striate Cortex*, J. Physiol., (1959), 148, 574-591
- [10] Krohn J., Beyleveld G., Bassens A., *Uczenie głębokie i sztuczna inteligencja. Interaktywny przewodnik ilustrowany*, Helion 2022
- [11] Kumar S., (2021) *7 Applications of Auto-Encoders every Data Scientist should know. Essential guide to Auto-Encoders and its usage*, Towards Data Science, <https://towardsdatascience.com/6-applications-of-auto-encoders-every-data-scientist-should-know-dc703cbc892b> (dostęp: 01.04.2022)
- [12] LeCun Y., Cortes C., Burges C. J. C., *THE MNIST DATABASE of handwritten digits* <http://yann.lecun.com/exdb/mnist/> (dostęp: 29.05.2022)

## Bibliografia

---

- [13] McCoy J. T., Kroon S., Auret L., *Variational Autoencoders for Missing Data Imputation with Application to a Simulated Milling Circuit*, IFAC-PapersOnLine, Volume 51, Issue 21, 2018, Pages 141-146, ISSN 2405-8963, <https://doi.org/10.1016/j.ifacol.2018.09.406>
- [14] Mungoli A., (2020) Dimensionality Reduction: PCA versus Autoencoders. Comparison of PCA and AutoEncoders for Dimensionality Reduction, Towards Data Science, <https://towardsdatascience.com/dimensionality-reduction-pca-versus-autoencoders-338fcf3297d> (dostęp: 24.04.2022)
- [15] Mishra D., *Transposed Convolution Demystified*, <https://towardsdatascience.com/transposed-convolution-demystified-84ca81b4baba> (dostęp: 15.05.2022)
- [16] Pinaya W. H. L., Vieira S., Garcia-Dias R., Mechelli A., *Machine Learning. Methods and Applications to Brain Disorders*, Academic Press, 2020
- [17] Programmatically, *An Introduction to Neural Network Loss Functions*, <https://programmatically.com/an-introduction-to-neural-network-loss-functions/> (dostęp: 15.05.2022)
- [18] Pröve P. L., *An Introduction to different Types of Convolutions in Deep Learning*, <https://towardsdatascience.com/types-of-convolutions-in-deep-learning-717013397f4d> (dostęp: 15.05.2022)
- [19] Skansi S. (2018) *Introduction to Deep Learning. From Logical Calculus to Artificial Intelligence*, Springer International Publishing
- [20] Susik, R. (2020). *Recurrent autoencoder with sequence-aware encoding* ArXiv. <https://doi.org/10.48550/ARXIV.2009.07349>
- [21] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P., *Image Quality Assessment: From Error Visibility to Structural Similarity*, IEEE Transactions on Image Processing, Vol. 13, No. 4, April 2004
- [22] Wong A., *Similar Image Retrieval using Autoencoders*, <https://towardsdatascience.com/find-similar-images-using-autoencoders-315f374029ea> (dostęp: 15.05.2022)

## Spis rysunków

1.1	Przykładowe sztuczne sieci neuronowe rozwiązuające proste zadania logiczne . . . . .	7
1.2	Struktura sztucznego neuronu, który stosuje funkcję skokową $f$ na ważonej sumie sygnałów wejściowych . . . . .	8
1.3	Perceptron z trzema neuronami wejściowymi i trzema wyjściami . . . . .	9
1.4	Przykładowe funkcje aktywacji wraz z pochodnymi . . . . .	11
1.5	Działanie neuronów biologicznych w korze wzrokowej . . . . .	13
1.6	Warstwy splotowe z prostokątnymi lokalnymi polami recepcyjnymi . . . . .	14
1.7	Związek pomiędzy warstwami a uzupełnianiem zerami . . . . .	14
1.8	Warstwa splotowa z krokiem o długości 2 . . . . .	15
1.9	Uzyskiwanie dwóch map cech za pomocą dwóch różnych filtrów . . . . .	16
1.10	Warstwy splotowe zawierające wiele map cech, a także zdjęcie z trzema kanałami barw .	17
1.11	Maksymalizująca warstwa łącząca (jądro łączące: $2 \times 2$ , krok: 2, brak uzupełniania zerami)	19
1.12	Warstwa <i>max-pooling</i> z filtrem i krokiem rozmiaru $2 \times 2$ , zastosowana do mapy aktywacji o rozmiarze $4 \times 4$ (widoczna po lewej stronie) . . . . .	19
1.13	Niezmienniczość związana z drobnymi przesunięciami . . . . .	20
1.14	Porównanie warstwy łączącej maksymalizującej i uśredniającej . . . . .	21
1.15	Struktura autoenkodera odwzorowującego wejście $x$ na wyjście $r$ (nazywane rekonstrukcją) poprzez reprezentację ukrytą (kodowanie) $h$ . Autoenkoder składa się z dwóch składników: kodera $f$ (odwzorowującego $x$ na $h$ ) i dekodera $g$ (odwzorowującego $h$ na $r$ ) . . . . .	22
1.16	Funkcje straty rzadkości . . . . .	26
1.17	Autokodery odszumiające: wykorzystujące szum gaussowski (po lewej) lub metodę porzucania (po prawej) . . . . .	27
1.18	Zaszumione obrazy (na górze) i ich rekonstrukcje (na dole) . . . . .	27
1.19	Struktura autoenkodera odszumiającego . . . . .	28
1.20	Przykładowa struktura autoenkodera stosowego . . . . .	29
1.21	Działanie warstwy ekspansji (nadpróbkowania) . . . . .	30
1.22	Działanie warstwy splotowej transponowanej . . . . .	31
1.23	Przykład autoenkodera rekurencyjnego . . . . .	33
1.24	Przykładowa struktura autoenkodera wariancyjnego . . . . .	34

## Spis rysunków

---

1.25 Reprezentacje ukryte. W tym przykładzie, obrazy cyfr od 0 do 9 były użyte do trenowania autoenkodera odszumiającego i przeciwnego. W obu przypadkach dane uczące są przedstawione w dwuwymiarowej przestrzeni ukrytej. Ponieważ nie nakładamy żadnych ograniczeń na autoenekoder odszumiający, to jego przestrzeń ukryta zawiera puste miejsca. Z drugiej strony, autoenekoder przeciwny ogranicza reprezentacje do podobnych według rozkładu (w tym przypadku dwuwymiarowego normalnego), co w rezultacie oznacza brak pustych miejsc . . . . .	36
1.26 Struktura generatywnej sieci przeciwniej (GAN) . . . . .	37
1.27 Struktura autokodera przeciwnego. Sieć dyskryminująca jest dodana do autoenkodera, aby zmusić go do generowania reprezentacji ukrytych podobnych do rozkładu a priori . . . . .	37
1.28 Autodenkoder w zastosowaniu do redukcji wymiarowości . . . . .	39
1.29 Autoenekoder stosowany do ekstrakcji cech . . . . .	40
1.30 Stosowanie autoenkodera do odszumiania obrazu . . . . .	41
1.31 Zastosowanie autoenkodera do kompresji obrazu . . . . .	42
1.32 Zastosowanie autoenkodera do wyszukiwania obrazu . . . . .	42
1.33 Wybór wartości progowej dla anomalii w autoenkoderze na podstawie funkcji straty . .	43
 2.1 Pięć pierwszych obserwacji ze zbioru cyfr MNIST . . . . .	45
2.2 Struktura autoenkodera prostego . . . . .	46
2.3 Reprezentacja 15-wymiarowa dla pierwszej obserwacji ze zbioru testowego MNIST . .	46
2.4 Pięć pierwszych obserwacji ze zbioru testowego MNIST: w górnym rzędzie oryginalne obrazy, w dolnym rekonstrukcje z autoenkodera niedopełnionego . . . . .	47
2.5 Wartość funkcji straty w kolejnych epokach trenowania autoenkodera niedopełnionego .	47
2.6 Struktura autoenkodera splotowego . . . . .	48
2.7 Pięć pierwszych obserwacji ze zbioru testowego MNIST: w górnym rzędzie oryginalne obrazy, w dolnym rekonstrukcje z autoenkodera splotowego . . . . .	49
2.8 Wartość funkcji straty w kolejnych epokach trenowania autoenkodera splotowego . . .	50
2.9 Zaszumione obrazy ze zbioru testowego MNIST . . . . .	50
2.10 Struktura autoenkodera odszumiającego . . . . .	51
2.11 Pięć pierwszych obserwacji ze zbioru testowego MNIST: w górnym rzędzie oryginalne obrazy, w środkowym zaszumione, w dolnym rekonstrukcje z autoenkodera odszumiającego . . . . .	51
2.12 Wartość funkcji straty w kolejnych epokach trenowania autoenkodera odszumiającego .	52
2.13 Zbiór testowy dla zadania wyszukiwania obrazu . . . . .	52
2.14 Schemat wyszukiwania podobnych obrazów za pomocą autoenkodera . . . . .	53
2.15 Wyniki zastosowania prostego autoenkodera przy wyszukiwaniu obrazu . . . . .	53
2.16 Struktura autoenkodera splotowego użytego do wykrywania anomalii . . . . .	54
2.17 Schemat wykrywania anomalii za pomocą autoenkodera . . . . .	55

2.18 Wartości funkcji straty z autoenkodera do wykrywania anomalii, z zaznaczoną wartością progową, powyżej której obserwacja jest uznawana za anomalię . . . . .	56
2.19 Przykłady anomalii w zbiorze MNIST wykrytych za pomocą autoenkodera . . . . .	56
2.20 Struktura autoenkodera wariancyjnego . . . . .	57
2.21 Dwa pierwsze wymiary przestrzeni ukrytej autoenkodera wariancyjnego . . . . .	58
2.22 Przykłady obrazów cyfr wygenerowanych przez autoenkoder wariancyjny . . . . .	59
2.23 Przestrzeń ukryta autoenkodera wariancyjnego z wymiarem kodowania ukrytego równym dwa . . . . .	60
2.24 Przestrzeń ukryta w postaci obrazów cyfr . . . . .	61



## **Spis tabel**

2.1	Struktura kodera w autoenkoderze prostym . . . . .	46
2.2	Struktura dekodera w autoenkoderze prostym . . . . .	46
2.3	Struktura kodera w autoenkoderze splotowym . . . . .	49
2.4	Struktura dekodera w autoenkoderze splotowym . . . . .	49
2.5	Struktura kodera w autoenkoderze splotowym użytym do wykrywania anomalii . . . . .	54
2.6	Struktura dekodera w autoenkoderze splotowym użytym do wykrywania anomalii . . . . .	55
2.7	Struktura kodera w autoenkoderze wariancyjnym . . . . .	57
2.8	Struktura dekodera w autoenkoderze wariancyjnym . . . . .	58



## **Załączniki**

1. Płyta CD z niniejszą pracą w wersji elektronicznej.



## **Streszczenie (Summary)**

### **Przegląd autoenkoderów stosowanych w nienadzorowanym uczeniu maszynowym**

Ta praca przedstawia pojęcia związane z sieciami neuronowymi, w szczególności ze splotowymi, takie jak: warstwy splotowe, łączące, filtry, mapy cech. Następnie wprowadzona jest definicja autoenkodera. Szeroko opisane są rodzaje autoenkoderów, używane w nich funkcje straty oraz ich zastosowania praktyczne. W części praktycznej przedstawione są przykłady użycia różnych rodzajów autoenkoderów na zbiorze danych MNIST, takie jak autoencoder prosty, splotowy, odszumiający, wariacyjny, a także zastosowania autoenkoderów do wykrywania anomalii oraz do wyszukiwania obrazów.

### *An overview of autoencoders used in unsupervised machine learning*

*This thesis presents concepts related to neural networks, in particular to convolutional networks, such as convolutional layers, pooling layers, filters, and feature maps. Then the definition of autoencoder is introduced. Types of autoencoders, loss functions used in them and their practical applications are described extensively. In the practical section, examples of using different types of autoencoders on the MNIST dataset are presented, such as simple autoencoder, convolutional autoencoder, denoising autoencoder, variational autoencoder, as well as applications of autoencoders for anomaly detection and image retrieval.*