



**WYDZIAŁ FIZYKI
i INFORMATYKI STOSOWANEJ**
Uniwersytet Łódzki

Maciej Pracucik

Kierunek: informatyka

Specjalność: informatyka stosowana

Ścieżka dydaktyczna: systemy mobilne

Numer albumu: 410731

Wykorzystanie sieci typu GAN na potrzeby generowania gestów rąk.

Praca magisterska

wykonana pod kierunkiem
dr Krzysztof Podlaski
w Katedrze Informatyki WFiIS UŁ

Łódź 2025

Spis treści

1 Wprowadzenie	4
1.1 Problematyka	4
1.2 Cel i zakres pracy	4
1.3 Struktura pracy	4
2 Sieci GAN, analiza konkurencji i techniczne aspekty realizacji	5
2.1 Sieci neuronowe i AI	5
2.2 Budowa sieci neuronowej	5
2.3 Sieci typu GAN	7
2.4 GestureGAN	8
2.5 PoseGAN	10
3 Narzędzia i technologie wybrane do realizacji projektu	12
3.1 Python	12
3.2 PyTorch	13
3.3 MediaPipe	13
3.4 OpenCV	13
4 Proces tworzenia projektu	14
4.1 Dobór zbioru danych	14
4.2 Wybór architektury sieci	14
4.3 Preprocesowanie danych	14
4.4 Tworzenie modelu	14
4.5 Tesotowanie modelu	14
4.6 Realizacja projektu	14
5 Wyniki i dyskusja	14
6 Podsumowanie	15
6.1 Zalety i wady przyjętych rozwiązań	15
6.2 Napotkane trudności	15
6.3 Możliwości rozwoju	15
6.4 Wnioski końcowe	15

Bibliografia 16

Listingi 19

1 Wprowadzenie

1.1 Problematyka

W dzisiejszych czasach bardzo modnym tematem jest sztuczna inteligencja, która zaczyna się wkradać w każdy aspekt naszego życia. Możemy ją spotkać w formie chatów, podpowiedzi do pisanego kodu, asystentów internetowych, systemów rozpoznawania głosów, czy nawet we własnej lodówce! Każda firma żeby zaistnieć i pozostać istotną inwestuje w tę część technologii. Jednakże to z czym AI radzi sobie najgorzej są ręce.

W tym projekcie chodzi o stworzenie sieci typu GAN, która pozwoli na generowanie obrazów rąk, w jak najlepszej jakości. Dodatkowo sieć ma za zadanie nauczyć się, żeby móc modyfikować istniejące zdjęcia i nadawać im zupełnie inny gest, przy zachowaniu jakości i realizmu. Szczególnie ten drugi aspekt pozostaje dla sztucznej inteligencji problematyczny. Myślę, że każdy z nas spotkał się ze zdjęciami, które dosłownie wyglądają, jak żywe, ale to co najczęściej zdradza, że jednak to AI maczało w nim palce są ręce. Za długie palce, dziwne ich ułożenie, ilość, czy nawet całkowicie odrealniony wygląd. Przeróżne firmy, jak i naukowcy stale ulepszają sieci, i rozwiązania, żeby i to przestało być problemem. Niniejsza praca również podejmuje się tego niełatwego zadania.

1.2 Cel i zakres pracy

Celem niniejszej pracy jest stworzenie sieci neuronowej typu GAN, która pozwoli na generowanie obrazów gestów rąk, w jak najlepszej jakości.

Docelowo również, wygenerowane zdjęcia będą wykorzystywane do stworzenia animacji przechodzenia z jednego gestu w inny.

1.3 Struktura pracy

Pierwszy rozdział przybliży to czym są sieci neuronowe, a dokładniej typu GAN, jakie są analogiczne rozwiązania, oraz o samym generowaniu zdjęć. Następny opowie jakie narzędzia, biblioteki i technologie zostały wykorzystane do realizacji projektu. Trzeci zaś mówi o tym jak wyglądał proces tworzenia projektu. Co po kolei zostało zrobione, jakie po drodze wystąpiły komplikacje, oraz jak zostały rozwiązane i finalnie jak wygląda projekt. Przedostatni rozdział to przedstawienie wyników, rezultatów realizowanego projektu, analiza i omówienie ich. Ostatni rozdział zawiera wnioski końcowe i podsumowanie.

2 Sieci GAN, analiza konkurencji i techniczne aspekty realizacji

2.1 Sieci neuronowe i AI

Człowiek od samego początku swojego istnienia jest istotą niesamowicie ciekawą. To ona sprawia, że w głowie ludzi pojawiają się pytania, a co jeśli? A co jeśli to co wiemy to jest tylko część prawdy? Co jeśli jest coś więcej? To dzięki zadawaniu sobie przeróżnych pytań przez różne osoby, najczęściej przez największe umysły jakie chodziły po tej ziemi tak dużo udało nam się osiągnąć. Poczynając od wynalezienia koła, silniki parowe, elektryczność, internet aż po loty w kosmos. Niektóre z tych wynalazków łączy kolejny aspekt, to że człowiek chce sobie ułatwiać codzienne zadania. Żeby jak najbardziej zwiększyć swoją produktywność, żeby codziennie czynności nie wchodziły w drogę, albo żeby je po prostu ułatwić czy przyspieszyć.

Kolejnym taki wynalazkiem, który łączy te dwa aspekty jest sztuczna inteligencja. Już od dawien dawna, przeróżne filmy czy książki science-fiction rozbudzały naszą wyobraźnię, o istnieniu robotów, sztucznej inteligencji równej ludzkiej, czy nawet przewyższającej ją. Do stworzenia pierwszych wzorów, inspiracją była ludzka sieć neuronowa, to jak działa nasz mózg. Takie coś starano się przekuć w pierwszej wzory, pierwsze sieci. Początkowo był to prosty matematyczny opis komórki neuronowej przez McCullocha i Pittsa w 1943 [6]. W połączeniu z zagadnieniem przetwarzania danych mógł modelować proste funkcje logiczne. Dopiero w 1949 roku Hebb sformułował regułę, którą uznaje się za pierwszą regułę uczenia sztucznych sieci neuronowych[6].

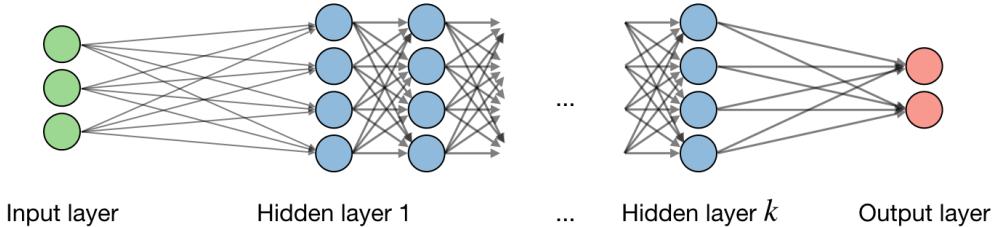
Obecnie sieci i modeli sieci jest tysiące. Do najpopularniejszych należą np CNN - Convolutional Neural Network, czy RNN - Recurrent Neural Network. W tym projekcie wykorzystujemy sieci typu GAN, czyli Generative Adversarial Network, o której więcej w kolejnym podrozdziale.

2.2 Budowa sieci neuronowej

Najmniejszą składową sieci neuronowej są oczywiście neurony, te neurony są poukładane po kilka w tak zwane warstwy. Najmniejsza ilość warstw to 3. Składają się z warstwy wejściowej, reprezentuje dane wejściowe w postaci numerycznej, warstwy ukryte, liczba mnoga tu jest celowa, ponieważ ich może wystąpić nieskończonie wiele, to one wykonują

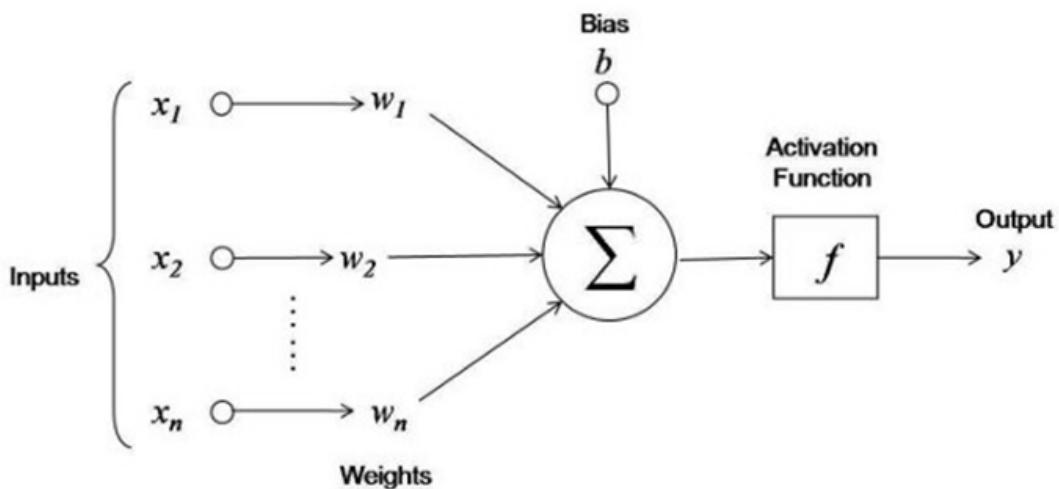
obliczenia. Ostatnim typem warstwy jest warstwa wyjściowa, ona generuje dane wyjściowe.

Tylko no właśnie dane. Jako, że sieci neuronowe to matematyka, to muszą działać na liczbach także nasze dane muszą być zmienione na liczby, a następnie znormalizowane do zakresu między 0 a 1.



Rysunek 1: Budowa sieci neuronowej [9]

Zajmijmy się teraz najmniejszą częścią sieci, czyli neuronem.



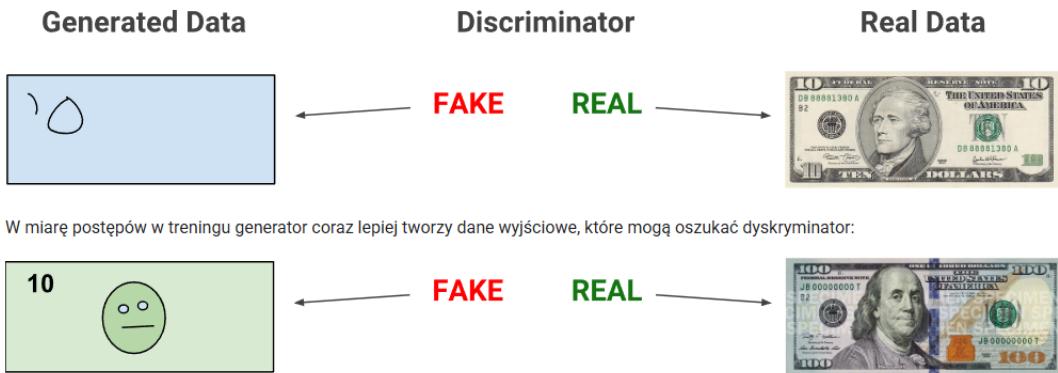
Rysunek 2: Budowa neuronu [9]

”W neuronie wartości sygnałów wejściowych (oznaczone na rysunku jako x_1, x_2, \dots, x_n) mnożone są przez wagi (oznaczone jako w_1, w_2, \dots, w_n), a iloczyny są ze sobą dodawane (do wyniku dodawana jest wartość b , która nie jest mnożona przez wartość sygnału wejściowego). Wynik rachunków poddawany jest funkcji aktywacji, która decyduje o ostatecznej wartości wysyłanego sygnału. W najprostszym przypadku może to być funkcja progowa (1 dla wartości dodatnich, 0 dla wartości niedodatnich), która przypomina działanie ludzkiego neuronu: „wysyłam sygnał lub nie”.” [9]

2.3 Sieci typu GAN

Sieci GAN czyli Generative Adversarial Network, a tłumacząc na polski, generatywne sieci współzawodnicze. Jak sama nazwa wskazuje są to sieci generatywne, czyli generują dane, najczęściej są to obrazy, ale nie ma ograniczeń, mogą być to również np filmiki. Sieci GAN składają się tak naprawdę z dwóch sieci, które ze sobą równocześnie współzawodnią. Bardzo często są porównywane do fałszowania pieniędzy, mamy jedną sieć, która się uczy generować jak najdokładniejsze fałszywki, a drugą, która ma za zadanie odrzucać fałszywki. Podzielone są na:

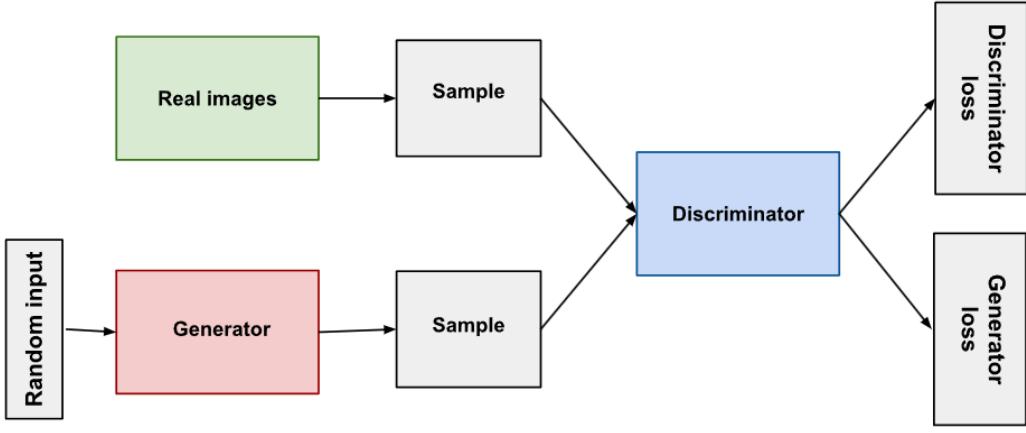
- Generator uczy się generować wiarygodne dane i docelowo oszukać dyskryminator, że to co generuje jest prawdziwe
- Dyskryminator, uczy się odróżniać prawdę od fikcji



Na koniec, jeśli szkolenie generatora przebiega dobrze, funkcja dyskryminacyjna coraz gorzej odróżnia prawdziwe obrazy od fałszywych. Zaczyna klasyfikować fałszywe dane jako prawdziwe, przez co jego dokładność spada.



Rysunek 3: Generative Adversarial Network [3]



Rysunek 4: Ogólny model sieci GAN [3]

Sieci GAN mają oczywiście swoje pod typy, czy też może framework, czy jeszcze inaczej już wcześniej stworzone typy tej sieci, które są powszechnie znane i używane. Wyróżnić można:

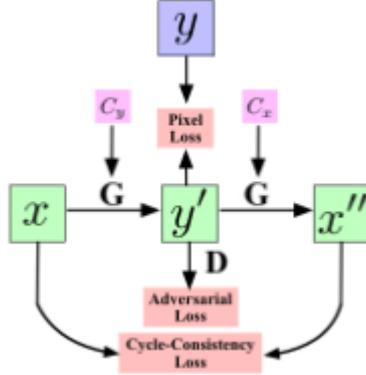
- Pix2Pix jest to sieć, która przetwarza jedno zdjęcie w drugie, co jest istotne w tej sieci to to, że dane muszą być sparowane, czyli jeden z obrazów jest obrazem wejściowym, a drugim obrazem wyjściowym
- CycleGAN jest bardzo podobny do Pix2Pix tylko zbiór danych nie jest sparowany. Można za jego pomocą robić takie translacje jak np. z koni zebry

Ten pierwszy został wykorzystany do stworzenia tego projektu, choć początkowo to ten drugi był używany, ale o tym w oddzielnym rozdziale.

2.4 GestureGAN

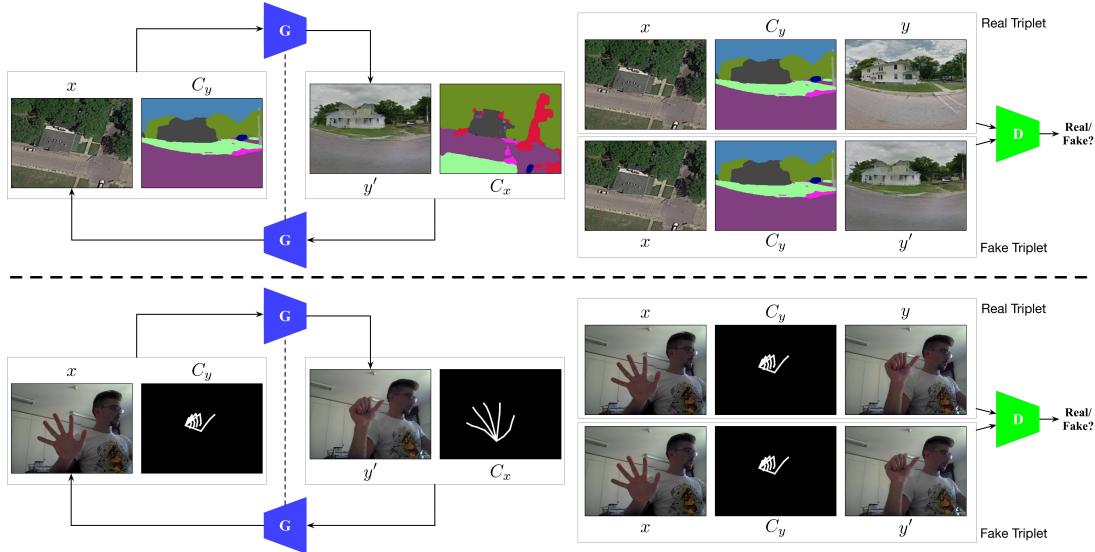
GestureGAN[5] jest to propozycja stworzona przez specjalistów z różnych uczelni, takich jak OXform czy Texas State University. W swoim założeniu ma robić to samo co sieć stworzona na potrzeby tego projektu, jednakże to jej budowa jest tym co je rozróżnia. Była to bardzo duża inspiracja przy tworzeniu projektu, podobnie jak niniejsza sieć wykorzystywany jest mediapipe do ekstrakcji szkieletu. Co je np. rozróżnia to, że GestureGAN otrzymuje dwa zdjęcia na wejściu, jedno prawdziwe, drugie samego szkieletu docelowego i ma wyjściu otrzymujemy wygenerowane zdjęcie i pierwotny szkielet. Takie rozwiązanie

było ok w przypadku ichniego zbioru danych, ponieważ był sparowany, to znaczy ta sama osoba wykonywała wszystkie gesty. Niestety w wykorzystanym przeze mnie zbiorze danych nie było takiej możliwości, dane stały się sparowane nieco sztucznie, ale o tym później. Bardzo dużą zaletą tego projektu jest to, że jest bardzo uniwersalny, z każdego gestu jesteśmy w stanie uzyskać każdy gest.



Rysunek 5: Model GestureGAN [5]

i tu jeszcze przykładowe efekty i możliwości GestureGAN[5].



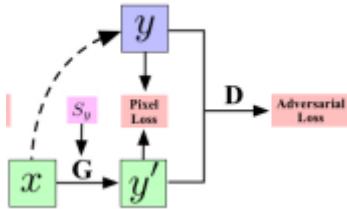
Rysunek 6: Zdjęcie wygenerowane przez GestureGAN [5]

Ten projekt stanowiłby bardzo dobrą bazę, generuje wyraźnie i zróżnicowane gesty rąk, jest bardzo uniwersalny, ale problem zaczyna się już na początku. Zbiór danych, który tutaj jest wykorzystywany jest niedostępny, albo wymaga dodatkowych dostępów. Nie wszystkie dane też są dostępne, choćby zdjęcia szkieletów. Te dane, które były dostępne,

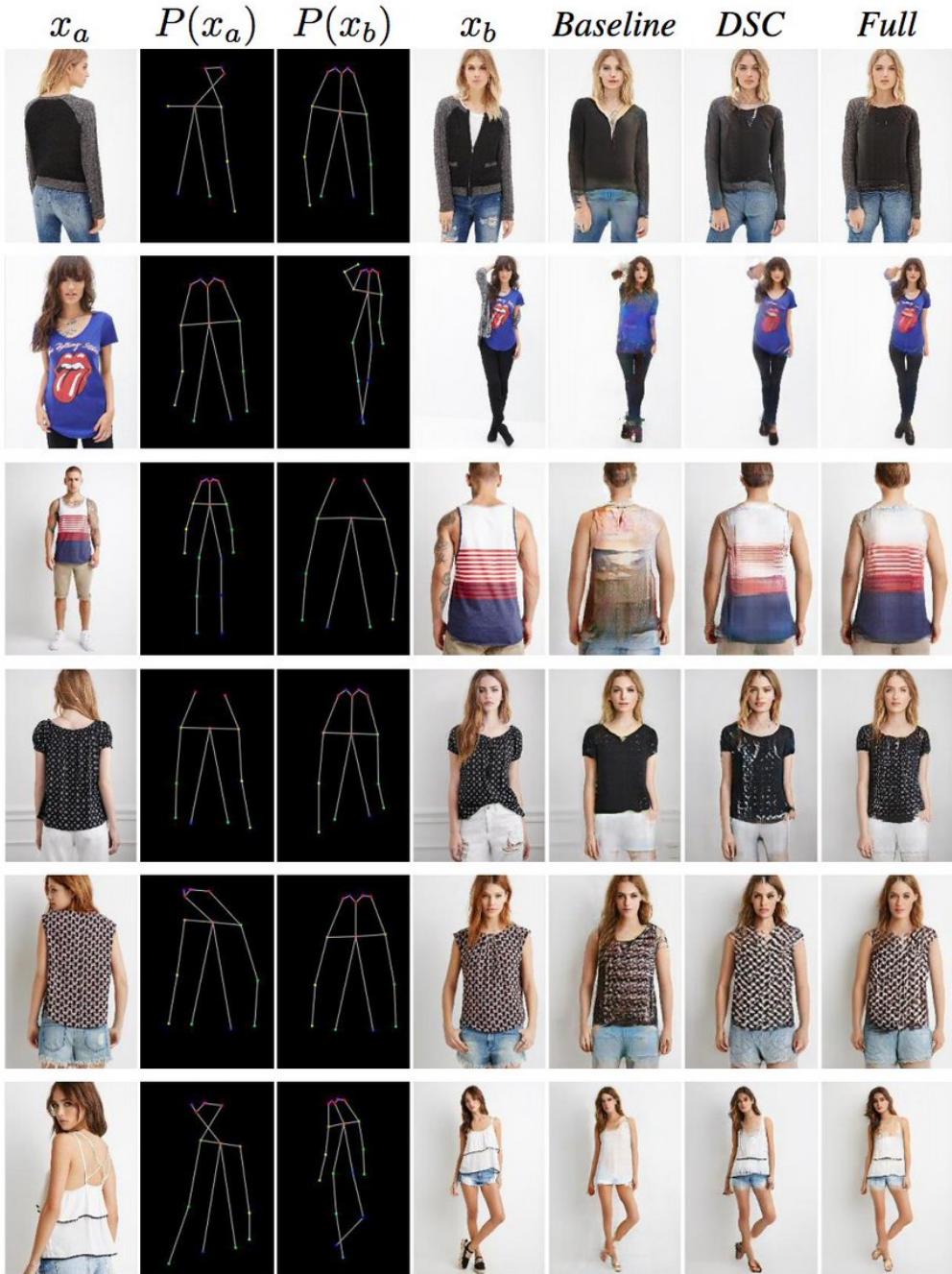
było ich po prostu za mało, plus możliwości tego projektu nie są aż tak potrzebne do realizacji tego, plus nie wiadomo jak z rozszerzalnością kodu. Dlatego najlepszym rozwiązaniem było przygotowanie własnego zbioru danych, własnego modelu i metodą prób i błędów, dochodzenie do rozwiązania.

2.5 PoseGAN

PoseGAN[1] jest dość zbliżony do GestureGAN, jednakże ma zasadniczą różnicę w swoim zastosowaniu. Tak jak GestureGAN, jak sama nazwa wskazuje służy do generowania gestów rąk, tak PoseGAN służy do generowania całych poz ludzkich. Także zakres jest większy jednakże przez fakt, że nie ma skupienia na dłoniach dokładność wykonywanego gestu nie ma tu znaczenia. Jednakże jest to zdecydowanie najbliższe wykonywanemu projektowi.



Rysunek 7: Model PoseGAN [1]



Rysunek 8: Zdjęcie wygenerowane przez PoseGAN [1]

Projekt jest niewątpliwie bardzo imponujący jednakże no właśnie skupia się na czymś kompletnie innym. Tak jak widać na 8 dlonie nawet nie są wyszczególnione na szkieletie, a przerabianie, a w zasadzie dorabianie takiego elementu byłoby zbyt czasochłonne. Sieć i tak by musiała się uczyć wszystkiego od nowa, plus no właśnie do końca interesuje nas cały ludzki szkielet. Dlatego niezbędne było stworzenie własnego projektu.

3 Narzędzia i technologie wybrane do realizacji projektu

3.1 Python

Językiem programowania wykorzystanym do realizacji projektu jest Python. Jest to bardzo oczywisty wybór, ponieważ jest to najpopularniejszy język programowania do tworzenia sieci neuronowych w dzisiejszych czasach. Przybliźmy go jednak, według oficjalnej dokumentacji, jest ”łatwy do nauczenia się i potężnym językiem programowania. Posiada wydajne struktury danych wysokiego poziomu oraz proste, ale skuteczne podejście do programowania obiektowego”[2]. Co niejako wyróżnia go na tle innych języków to fakt, że jest dynamicznie typowany oraz jest językiem interpretowanym. To oznacza, że nie posiada kompilatora a interpreter, który nie kompiluje programu do pliku wykonywalnego, a kod jest wykonywany w czasie rzeczywistym. Czyni to Python językiem łatwym w testowaniu, kompilowaniu czy wykonywaniu, przenośny, co oznacza, że ten sam kod uruchomi się niezależnie od systemu operacyjnego czy urządzenia. Sam język Python jest zbudowany na bibliotece C, co oznacza, że jest bardzo szybki i wydajny. Został stworzony przez Guido van Rossum w roku 1991. Co ciekawe jego nazwa wywodzi się z starych serii skoczów grupy Monty Python’s Flying Circus, a zyskał na popularności w momencie gdy firma Google, powiedziała, że wykorzystuje go do własnych, wewnętrznych celów.

Jednakże czemu akurat to on jest najczęściej wykorzystany do AI? Odpowiedź jest tak prosta jak sam Python jest prosty. Wynika to z tego, że Python ma bardzo prostą i czytelną składnię, co pozwala developerom skupianie się na logice i samym problemie, a nie na składni[10]. Dodatkowo Python sam w sobie nie wymaga dużej ilości kodu. Najprostsza sieć neuronowa może zostać stworzona i uruchomiona w zaledwie 4 linijkach! Kolejnym powodem przemawiającym dlaczego to właśnie Python jest najczęściej używany, jest bardzo duża ilość bibliotek, czy to wbudowanych, czy stworzonych przez społeczność. Biblioteki takie jak NumPy, SciPy, Matplotlib, czy wykorzystane w tym projekcie PyTorch czy MediaPipe, o których będzie w kolejnych podrozdziałach. Także kolejnym i ostatnim już aspektem, o którym chcę wspomnieć, jest wyżej wymieniona społeczność. To dzięki dużej liczbie osób i ogromnym zebranym doświadczeniu, tworzenie dowolnego projektu staje się znacznie prostsze. Praktycznie każdy projekt, aspekt projektu czy problem natrafiało wcześniej duża część ludzi, dzięki czemu możemy szybko i sprawnie rozwiązywać pro-

blem.

3.2 PyTorch

”PyTorch to w pełni funkcjonalny framework do tworzenia modeli do deep learningu, który jest typem machine learningu, najczęściej wykorzystywanym w aplikacjach takich jak rozpoznawanie obrazów czy procesowanie języka.”[7] Biblioteka ta została napisana w Pythonie, przez developerów z Facebook AI Research, oraz ma doskonałe wsparcie do wykorzystywania GPU, szczególnie dla GPU od firmy Nvidia, który posiadam. Co czyni ją idealną biblioteką dla tego projektu. Jednak projekt dotyczy generowania obrazów rąk, czyli dobre wykorzystanie GPU jest bardzo wskazane. Początkowo jednak używana była konkurencyjna biblioteka, a dokładniej Tensorflow, jednakże była zdecydowanie wolniejsza i mniej klarowna od PyTorch co zaważyło na finalnym wyborze.

3.3 MediaPipe

”MediaPipe Solutions to zestaw bibliotek i narzędzi, które umożliwiają szybkie stosowanie w aplikacjach technik sztucznej inteligencji (AI) i uczenia maszynowego (ML).”[4] Opis ten pochodzi z oficjalnej dokumentacji Mediapipe, jednakże sam opis jest zbyt ogólny. Ta biblioteka ma masę możliwości i zastosować. Do nich można zaliczyć np. rozpoznawanie obrazu, nie chcemy pisać sieci do rozpoznawania, czy na danym obrazku jest pies czy kot? Mediapipe daje gotowe rozwiązanie! Poza tym do wyboru jest też klasyfikacja obrazu, segmentacja obrazu, wykrywanie twarzy, rozpoznawanie gestu, oraz to co było niezbędne w tym projekcie to wykrywanie rąk. To dzięki tej bibliotece udało się wyłuskać szkielet ręki na każdym z obrazów ze zbioru danych, nałożyć go na zdjęcie wraz z maską i przekazane do modelu. Później te informacje były użyte do tworzenia kolejnych klatek animacji przechodzenia z jednego gestu w drugi.

3.4 OpenCV

OpenCV jest to biblioteka do machine learningu oraz computer vision, posiada w swoim arsenale dostęp do takich narzędzi jak wykrywanie i rozpoznawanie twarzy, identyfikowanie obiektów, śledzenie ruchów kamery, śledzenie obiektów 3D, usuwanie czerwonych oczu ze zdjęć, czy nawet łączenie zdjęć razem, by uzyskać obraz całej sceny o wysokiej

rozdzielczości.[8] Jednak to żadna z tych funkcji nie została użyta w projekcie. Wykrywanie i tworzenie szkieletu to zadanie Mediapipe, tworzenie modelu to działka PyTorch. OpenCV miał znacznie prostsze zadanie, zostało wykorzystane do ładowania zdjęć, czy to dla preprocessingu, czy już bezpośrednio do modelu. Dalej zapisywał każdy kolejny obraz w zależności od epoki, dzięki czemu można było śledzić poczynania i na koniec sklejał wszystkie klatki i tworzył z nich animację.

4 Proces tworzenia projektu

4.1 Dobór zbioru danych

zbior

4.2 Wybór architektury sieci

architektura

4.3 Preprocesowanie danych

Preprocesowanie

4.4 Tworzenie modelu

model

4.5 Tesotowanie modelu

Tesotowanie

4.6 Realizacja projektu

realizacja

5 Wyniki i dyskusja

wyniki

6 Podsumowanie

6.1 Zalety i wady przyjętych rozwiązań

Zalety

6.2 Napotkane trudności

trudności

6.3 Możliwości rozwoju

rozwoj

6.4 Wnioski końcowe

wnioski

Bibliografia

- [1] Aliaksandr Siarohin i Enver Sanginetto i Stephane Lathuiliere i Nicu Sebe. “Deformable GANs for Pose-based Human Image Generationn”. In: (2018).
- [2] Python Software Foundation. *The Python Tutorial*. 2025. URL: <https://docs.python.org/3/tutorial/index.html>.
- [3] Google. *Kurs zaawansowany Machine Learning GAN*. URL: <https://developers.google.com/machine-learning/gan?hl=pl>.
- [4] Google. *Przewodnik po rozwiązańach MediaPipe*. URL: <https://ai.google.dev/edge/mediapipe/solutions/guide?hl=pl>.
- [5] Hao Tang i Hong Liu i Nicu Sebe. “Unified Generative Adversarial Networks for Controllable Image-to-Image Translation”. In: (2020).
- [6] Bohdan Macukow. “SIECI NEURONOWE, HISTORIA BADAŃ I PODSTAWOWE MODELE”. In: (). URL: <https://pagesmini.edu.pl/~macukowb/wspolne/PNEiTI.pdf>.
- [7] Nvidia. *PyTorch*. URL: <https://www.nvidia.com/en-us/glossary/pytorch/>.
- [8] OpenCV. *OpenCV is the world’s biggest computer vision library*. URL: <https://opencv.org/about/>.
- [9] Blog CSI Uniwersytet Im. Adama Mickiewicza w Poznaniu. *Sieci neuronowe w NLP*. 2025. URL: <https://csi.amu.edu.pl/blog-csi/sieci-neuronowe-w-nlp>.
- [10] Dhruvitkumar Talati. “Python: The alchemist behind AI’s intelligent evolution”. In: (2021).

Spis rysunków

1	Budowa sieci neuronowej [9]	6
2	Budowa neuronu [9]	6
3	Generative Adversarial Network [3]	7
4	Ogólny model sieci GAN [3]	8
5	Model GestureGAN [5]	9
6	Zdjęcie wygenerowane przez GestureGAN [5]	9
7	Model PoseGAN [1]	10
8	Zdjęcie wygenerowane przez PoseGAN [1]	11

Spis tabel

Listingi