POLITECHNIKA WROCŁAWSKA
WYDZIAŁ Informatyki i Telekomunikacji

Głębokie sieci neuronowe

Rozpoznawanie wieku, płci oraz rasy człowieka na podstawie zdjęć twarzy

Kierunek: Informatyczne Systemy Automatyki

Prowadzący:
Dr hab. inż. Andrzej Rusiecki

Autorzy:
Filip Wojakiewicz
Przemysław Marciniak 247331

Wrocław 2023

Cel i zakres projektu

Głównym celem projektu jest implementacja rozwiązania rozpoznającego wiek, płeć oraz rasę przy wykorzystaniu zdjęć twarzy. Wyuczenie sieci neuronowej na podstawie zbioru uczącego do realizowania zadania rozpoznawczego. Dopasowanie parametrów sieci do otrzymywanych wyników. Analiza wyników oraz przeprowadzenie szeregu testów mających na celu sprawdzenie poprawności działania sieci.

Założenia projektowe

Zbiór danych został pobrany ze strony https://susanqq.github.io/UTKFace/. Zawiera on ponad 23 000 zdjęć twarz ludzi w przedziale wieku od 1 do 116 lat, o rozdzielczości 200x200 i przestrzeni barw RGB. Wszystkie zdjęcia zamieszczone w bazie danych zostały pozyskane z internetu dlatego duża część osób znajdująca się w bazie danych to znane osoby m.in. aktorzy, piosenkarze i celebryci. Każdy plik nazwany jest według schematu [wiek]_[płeć]_[rasa]_[data].jpg. Liczba w miejscu [wiek] bezpośrednio określa wiek osoby, w miejscu [płeć] znajduje się 0- określające mężczyznę lub 1- określające kobietę, natomiast [rasa] przyjmuje wartości od 0 do 4, które oznaczają kolejno rasę białą, czarną, Azjatę, Hindusa i inną (np. Latynos, Filipińczyk itp.). Cyfry na ostatnim miejscu oznaczają dokładną datę, w której zdjęcie trafiło do zbioru danych. Cały zbiór danych został publicznie udostępniony do niekomercyjnych badań naukowych.

W projekcie wykorzystane zostały m.in. następujące biblioteki:

- sklearn- wykorzystano m.in. train_test_split do podziału całego zbioru na zbiór treningowy i testowy, metrics.accuracy_score obliczający skuteczność danego modelu
- tensorflow- wykorzystano m.in. keras.losses.SparseCategoricalCrossentropy dzięki któremu uzyskuje się procentową przynależność danej osoby do kategorii (wykorzystywana jest przy modelu rasy), keras.callbacks.EarlyStopping powoduje zatrzymanie procesu uczenia gdy metryka przestanie się polepszać
- opencv- wykorzystano m.in. imread służący do wczytania zdjęcia z konkretnego pliku, cvtColor powodujący konwersje obrazu z jednej przestrzeni kolorów do drugiej (np. BGR na RGB)

Ze względu na duży przedział wiekowy wyuczone modele uzyskiwały niską skuteczność np. w przypadku osób młodych (szczególnie niemowląt) i bardzo starych (w wieku 90+) błędnie klasyfikowano płeć osoby. Z tego względu postanowiono podzielić cały zbiór według wieku na 3 podzbiory: 1-15, 15-55, 55-120. W przypadku modeli wieku uzyskiwano najmniejszą skuteczność (na poziomie 6%). Było to spowodowane tym, że zdjęcie było zaliczone do poprawnie sklasyfikowanych tylko w momencie, gdy program podał dokładny wiek osoby, co jest bardzo trudne nawet dla przeciętnego człowieka. Z tego powodu wprowadzony został przedział tolerancji wieku, w którym zdjęcie będzie uznawane za poprawnie sklasyfikowane. Dla przedziału 1-15 ustawiono ten przedział na +/-3, a dla pozostałych na +/-5.

Wykorzystane modele

Do realizacji problemu klasyfikacji człowieka po zdjęciu twarzy stworzono 3 niezależne modele: age_model (dla wieku), gender_model (dla płci) oraz race_model (dla rasy). Wszystkie modele są typu Sequential(), a więc każda warstwa ma dokładnie jeden tensor wejściowy i jeden tensor wyjściowy. Następnie, przy pomocy model.add(), dodawane są kolejne warstwy konwolucyjne (Conv2D) i poolingu (MaxPool2D). Liczba warstw dla poszczególnych modeli: age_model- 4 warstwy, gender_model- 5 warstw, race_model- 5 warstw. Wszystkie warstwy konwolucyjne zostały ustawione następująco: rozmiar filtra (kernel size) na 3, funkcja aktywacji to ReLU, warstwa wejściowa (input shape) o wymiarach 200x200 typu RGB. Warstwy różnią się od siebie liczbą filtrów wyjściowych w splocie.

Wartości te zostały dobrane w sposób eksperymentalny, metodą prób i błędów. W ten sposób otrzymano, dla kolejnych warstw, następującą ilość filtrów:

- age_model: 128, 128, 256, 512
- gender_model: 36, 64, 128, 256, 512
- race_model: 32, 64, 128, 256, 512

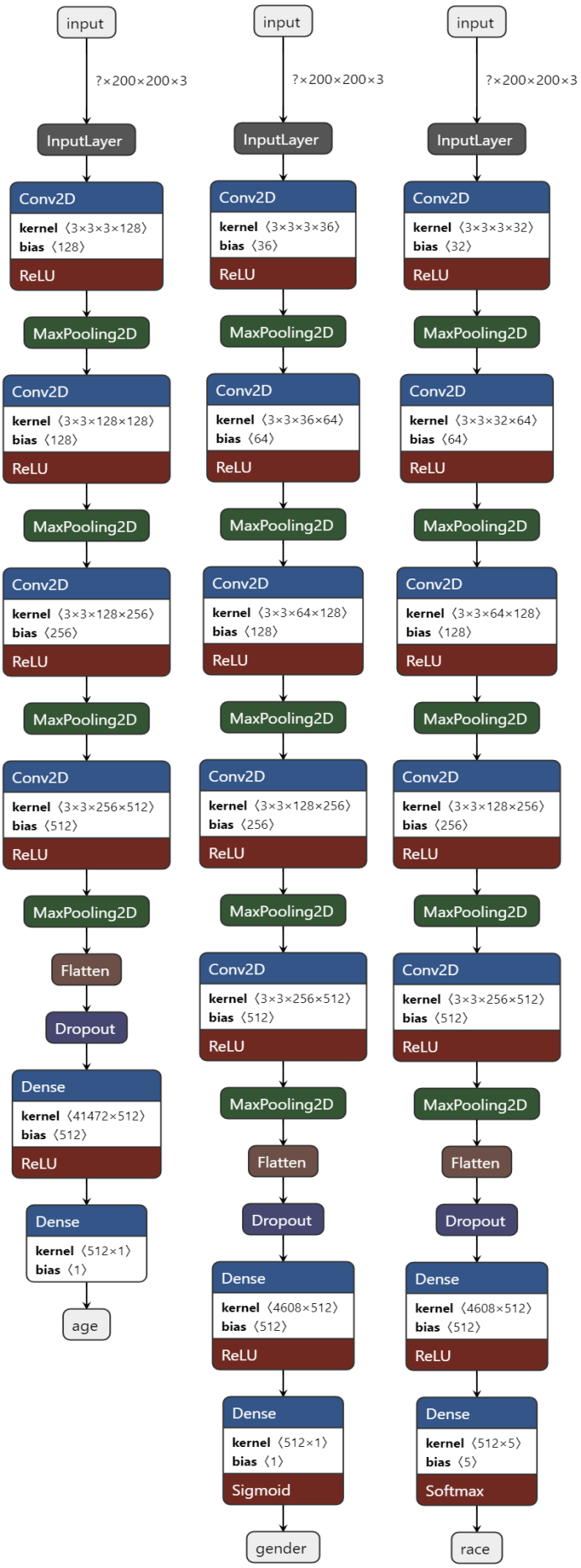
Wszystkie warstwy poolingu mają okno o wymiarach 3x3, a krok ustawiony na 2. Następną warstwą występującą w modelu jest warstwa spłaszczająca (flatten), której zadaniem jest przekształcenie 2-wymiarowej warstwy w 1-wymiarowy wektor. Warstwa porzucenia (dropout) ma za zadanie zapobiegać przetrenowaniu sieci. W każdym modelu wartość dropout została ustawiona na 20%. Ostatnimi warstwami w modelu są 2 warstwy gęste (dense layer). Wymiarowość przestrzeni wyjściowej warstwy 1 równa się ilości filtrów wyjściowych w ostatniej warstwie konwolucyjnej (w naszym przypadku jest to 512 dla wszystkich modeli), a funkcja aktywacji to ReLU. Druga warstwa gęsta przyjmuje następujące parametry w zależności od modelu:

- age_model: przestrzeń wyjściowa = 1, funkcja aktywacji = liniowa
- gender_model: przestrzeń wyjściowa = 1, funkcja aktywacji = sigmoid
- race_model: przestrzeń wyjściowa = 5, funkcja aktywacji = softmax

Parametry kompilacji poszczególnych modeli:

- age_model: optymalizator = adam, funkcja straty = MSE, miara = MAE
- gender_model: optymalizator = adam, funkcja straty = binary_crossentropy, miara = accuracy
- race_model: optymalizator = adam, funkcja straty = SparseCategoricalCrossentropy, miara = accuracy

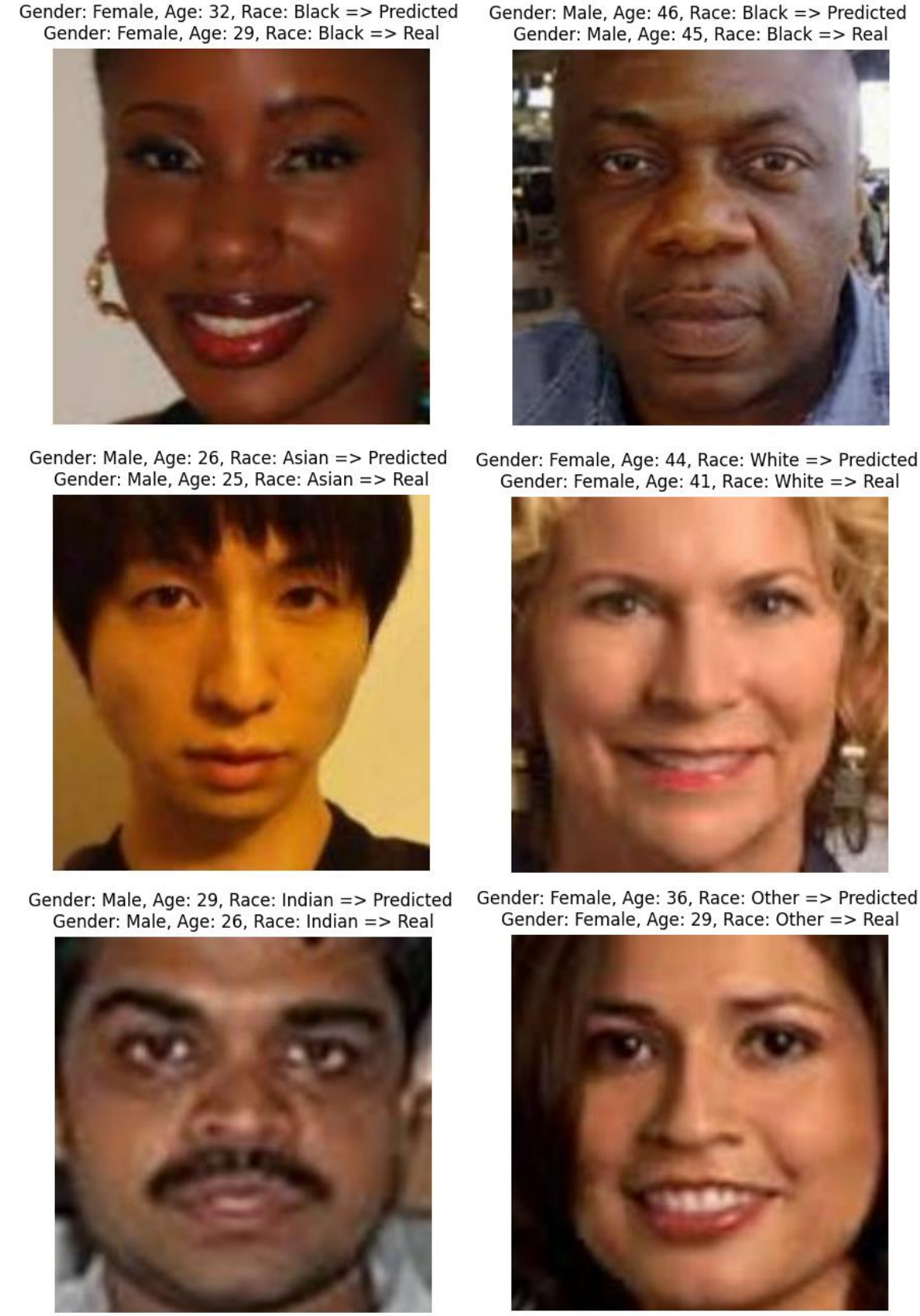
Schematy graficzne modeli przedstawiono na rys. 1.



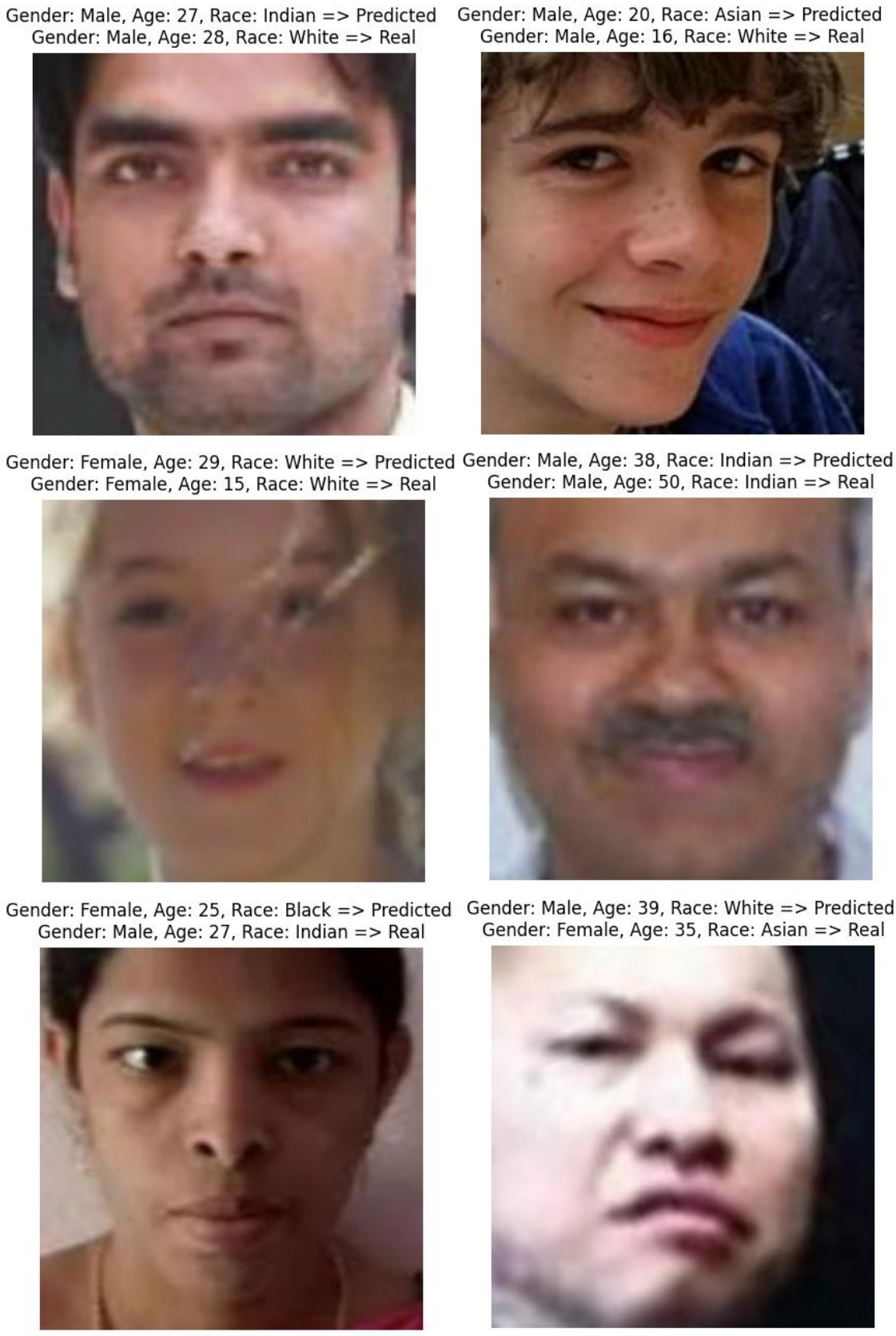
Rysunek 1 Schematy modeli sieci neuronowej (od lewej): age_model, gender_model, race_model

Wyniki

Rezultat pracy wyuczonych modeli dla przedziału wiekowego 15-55 lat zaprezentowano na rys. 2 i 3. Rys. 2 zawiera wyłącznie obrazy sklasyfikowane poprawnie, natomiast rys. 3, obrazy rozpoznane błędnie ze względu na jedną z cech.



Rysunek 2 Poprawnie sklasyfikowane twarze ludzi. Napis na górze informuje o wartościach przewidzianych przez program (Predicted), natomiast napis poniżej o rzeczywistych danych osoby (Real)



Rysunek 3 Niepoprawnie sklasyfikowane twarze ludzi. Napis na górze informuje o wartościach przewidzianych przez program (Predicted), natomiast napis poniżej o rzeczywistych danych osoby (Real)

Porównując ze sobą zdjęcia ludzi z rys. 2 i 3 można zauważyć, że znacząco różnią się one pod względem jakości, oświetlenia i widoczności całej twarzy. Rys. 3 we wszystkich tych

względach przedstawia zdjęcia dużo gorsze od tych zaprezentowanych w rys. 2. Wszystko to wpływa na błędną klasyfikację osób przez program. W ostatnich dwóch zdjęciach na rys. 3 gdzie źle sklasyfikowano płeć bardzo ciężko stwierdzić czy osoba na zdjęciu jest kobietą czy mężczyzną, z kolei błędna klasyfikacja rasy (szczególnie w przypadku zdjęcia w prawym dolnym rogu) może wynikać ze złego oświetlenia lub złego kontrastu. Błędy w predykcji wieku spowodowane są (oprócz, wcześniej wspomnianej dużej trudności odgadnięcia prawdziwego wieku osoby przez przeciętnego człowieka) specyfiką zbioru danych. Zdjęcia pochodzą z różnych stron internetowych i zawierają wielu popularnych celebrytów często wyglądających dużo młodziej niż faktycznie mają lat. Skutkiem tego jest częste zawyżanie prawdziwego wieku osoby młodej. Kolejnym problemem była znacząca przewaga próbek zawierających osoby w wieku 26 lat (ponad 10% całego zbioru). Z tego względu wprowadzono losowe pomijanie obrazów w tej kategorii wiekowej, ograniczając jej liczebność mniej więcej o połowę. Przyczyniło się to do zwiększenia skuteczności modelu.

Otrzymana skuteczność dla modeli w poszczególnych przedziałach wiekowych przedstawiono w Tab. 1

Skuteczność			
Przedział [lata]	1-15	15-55	55-120
Age	79%	53%	31%
Race	73%	76%	74%
Gender	64%	94%	84%

Tabela 1 Skuteczność modeli age, race i gender dla przedziałów 1-15 lat, 15-55 lat, 55-120 lat.

Największa skuteczność dla wieku obserwujemy w przedziale 1-15 lat. Pod względem lat jest to najmniejszy zbiór, więc program miał największą szansę podać wiek osoby w granicy tolerancji (która dla tego przedziału jest zmniejszona do +/-3). Niemowlęta bardzo łatwo odróżnić od dzieci w wieku szkolnym i od nastolatków, dlatego tutaj obserwujemy największą skuteczność ze wszystkich przedziałów. Grupa wiekowa 55-120 osiągnęła najniższy wynik- 31%. Wynika to z faktu, że ciężko jest precyzyjnie określić wiek starszej osoby, oraz że liczebność populacji niektórych lat była niewielka w porównaniu do pozostałych. W przypadku rasy, skuteczność we wszystkich modelach jest bardzo podobna. Dzięki wykorzystaniu SparseCategoricalCrossentropy uzyskujemy procentową przynależność osoby do każdej rasy i w efekcie uzyskujemy dużo lepszy wynik pracy modelu (wcześniej wykorzystywano wartości ułamkowe i zaokrąglano do najbliższej rasy, wtedy skuteczność była na poziomie 30%). Klasyfikacja płci najlepiej wypadła w przedziale wiekowym 15-55. Wtedy najłatwiej określić płeć osoby. Młode dzieci nie mają widocznych tak bardzo cech charakterystycznych dla poszczególnej płci i łatwo można się pomylić. W przypadku starszych osób, niektóre zdjęcia przedstawiają twarze z wieloma zmarszczkami które utrudniają ekstrakcje cech, stąd skuteczność wyniosła 84%.

Podsumowanie

Wykorzystując głębokie sieci neuronowe możliwe jest stworzenie programu rozpoznającego płeć, wiek oraz rasę człowieka tylko na podstawię zdjęcia twarzy. Aby tego dokonać stworzono trzy wielowarstwowe modele i przetestowano je pod kątem doboru najlepszych parametrów. Niska skuteczność w niektórych przypadkach wynika z samej złożoności problemu klasyfikacji, która, nawet dla człowieka, niekiedy jest bardzo trudna. W celu jej poprawienia należałoby znaleźć lub stworzyć lepszy zbiór danych zawierający portrety w większej rozdzielczości, w dobrym oświetleniu, z pełnym widokiem na twarz, a najlepiej w tej samej pozycji głowy (np. frontalnie do kamery). Ze względu jednak na ochronę danych osobowych stworzenie równie licznej i różnorodnej bazy danych co UTKFace jest bardzo trudne. Zbiór zdjęć musiał zostać odpowiednio przygotowany, podzielony na poszczególne kategorie wiekowe oraz ograniczono liczebność jednej klasy wiekowej.