Zeszyty Naukowe WWSI, No 23, Vol. 14 2020, s. 7-26
DOI: 10.26348/znwwsi.23.7

Opracowanie koncepcji i implementacja modelu rozpoznawania obrazu z wykorzystaniem elementów sztucznej inteligencji

Anna Sierżantowicz*, Andrzej Ptasznik**

Warszawska Wyższa Szkoła Informatyki

Streszczenie

W niniejszym artykule przedstawiono koncepcję i implementację modelu do rozpoznawania ras psów na podstawie zdjęcia. Do realizacji zadania wykorzystano model głębokiej sieci neuronowej bazujący na strukturze InceptionV3. Sieć została wytrenowana i przetestowana na zbiorze przypadków uczących liczącym ponad 20 tys. zdjęć 120 ras psów z zastosowaniem transferu wiedzy. Zbadano również wpływ jakości zdjęć na wyniki klasyfikacji. Sieć uzyskała bardzo dobre rezultaty zarówno w przypadku analizy typowych, jak i nietypowych zdjęć.

Słowa kluczowe **–** rozpoznawanie obrazu, sztuczne sieci neuronowe, sieci splotowe, InceptionV3, transfer wiedzy

* E-mail: anna.sierzantowicz@gmail.com
** E-mail: aptaszni@wwsi.edu.pl

Zgłoszono do druku 23 września 2020 r.

1. Wprowadzenie

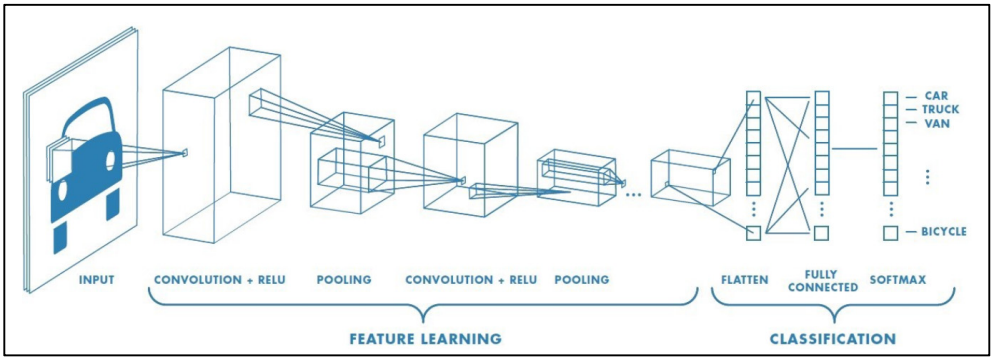
Tradycyjne programowanie dedykowane jest głównie do realizacji zagadnień, dla których istnieje możliwość zdefiniowania algorytmu postępowania. Niestety istnieje wiele złożonych problemów, takich jak: prognozowanie, aproksymacja funkcji, klasyfikacja czy analiza mowy i obrazu, dla których określenie algorytmu jest niemożliwe. Do realizacji tego typu zadań wykorzystywane są Sztuczne Sieci Neuronowe (SSN) – systemy przetwarzające informacje, wzorowane na budowie biologicznych układów nerwowych i zachodzących w nich procesach. Prace nad budową SSN rozpoczęto w ubiegłym stuleciu. Początkowo sztuczne sieci neuronowe realizowane były głównie w formie sprzętowej. Przykładem może być zbudowany przez Franka Rosenblatta i Charlesa Wightmana w roku 1957 perceptron do rozpoznawania znaków alfanumerycznych, będący pierwszą powszechnie znaną siecią neuronową [1]. Obecnie sztuczne sieci neuronowe realizowane są dość powszechnie również w formie programów komputerowych symulujących działanie sieci.

Współcześnie sztuczne sieci neuronowe znalazły zastosowanie w wielu dziedzinach naukowych. Bardzo dobrze sprawdzają się w zadaniach z zakresu rozpoznawania obrazu. Poza identyfikacją obiektów, potrafią również określać emocje bazując na mimice ludzkiej twarzy, identyfikować osoby publiczne, a także są w stanie analizować treści i klasyfikować je jako potencjalnie niebezpieczne. Od roku 2012 w zagadnieniach analizy obrazu coraz częściej stosowane są głębokie, splotowe sieci neuronowe (ang. *Convolution Neural Network*, CNN) [2]. Architektura tych sieci zasadniczo różni się od sieci klasycznych. Wartość funkcji aktywacji neuronu zależy od neuronów należących do pola recepcyjnego w warstwie poprzedniej, co znacząco ogranicza liczbę parametrów sieci. Dzięki temu proces uczenia przebiega dużo szybciej.

Celem pracy było zbadanie możliwości sieci neuronowych w zakresie rozpoznawania obrazu. W związku z powyższym, został zbudowany model do rozpoznawania ras psów na podstawie zdjęcia [3]. Ze względu na wymienione wyżej liczne zalety, do realizacji zadania została wykorzystana architektura splotowych sieci neuronowych. Zbudowany i wytrenowany model osiągnął bardzo dobre wyniki w fazie testów. W przypadku zdjęć nietypowych, model również osiągnął dobre rezultaty.

2. Splotowe sieci neuronowe

Splotowe sieci neuronowe mają bardzo specyficzną budowę. W ich architekturze można wyróżnić dwie charakterystyczne części, z których każda jest zbudowana z innego rodzaju warstw (rysunek 1). Pierwsza z nich jest odpowiedzialna za ekstrakcję cech i składa się z sekwencji warstw splotowych (ang. *convolution layers*) oraz warstw redukujących rozmiar (ang. *pooling layers*). Druga część jest dedykowana do zagadnień klasyfikacji i jest wzorowana na klasycznych sieciach neuronowych w pełni ze sobą połączonych (połączenia typu „każdy z każdym”).



Rysunek 1. Przykładowa architektura sieci splotowej [4]

Obraz dostarczany na wejście splotowej sieci neuronowej podlega wielokrotnym operacjom splotu, w wyniku których powstają mapy cech, zwane również mapami aktywacji. Do wykonania operacji splotu niezbędne jest zdefiniowanie filtra – macierzy o rozmiarach np. 3×3 – oraz skoku. Filtr przemieszcza się względem obrazu o zadany skok. Za każdym razem wykonywane jest mnożenie elementów filtra przez odpowiadające mu aktualnie elementy obrazu (lub w dalszych warstwach, elementy wejścia poprzedzającej warstwy). W przypadku obrazów kolorowych, obraz dzielony jest na 3 kanały, z których każdy przedstawia macierz dla jednej ze składowych modelu barwnego RGB. Operacje splotu są wykonywane osobno dla każdego kanału, po czym uzyskane wartości są sumowane celem utworzenia macierzy wartości wyjściowych warstwy.

W efekcie wykonywania operacji splotu, wartość funkcji aktywacji neuronów danej warstwy jest zależna od niewielkiej grupy neuronów warstwy poprzedniej, w przeciwieństwie do sieci klasycznych, w których wartość ta jest zależna od wszystkich neuronów warstwy poprzedzającej. Dodatkowo zachodzi zjawisko współdzielenia wag. Ze względu na fakt, iż w przypadku niektórych funkcji aktywacji (np. tangens hiperboliczny), dość szybko dochodzi do nasycenia funkcji i tym samym do zanikania gradientu, w sieciach splotowych zaczęto stosować funkcję aktywacji ReLU (ang. *Rectified Linear Units*) [2]. Funkcja ta określona jest wzorem: $f(x) = x^+ = \max(0, x)$. Badania wykazały, że stosowanie funkcji aktywacji ReLU w sieciach splotowych zmniejsza złożoność obliczeniową oraz przyspiesza proces uczenia [2].

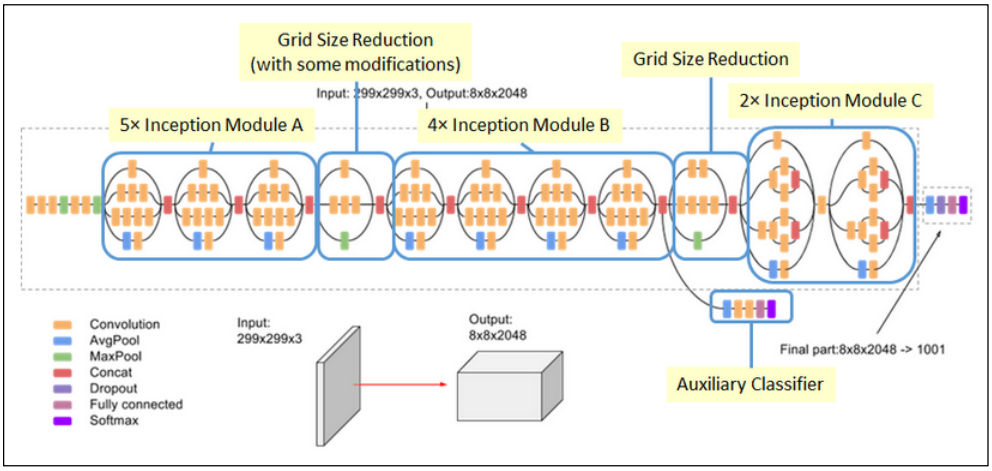
Kolejnym zabiegiem redukującym rozmiar macierzy wyjściowej, a tym samym liczbę parametrów, jest okresowe dodawanie warstwy grupowania (ang. *pooling layer*). W warstwie tej, wartości z pola recepcyjnego są agregowane. Zazwyczaj do agregacji stosuje się funkcję maksimum bądź średnią arytmetyczną. W wyniku tych działań eksponowane są wartości oznaczające obecność danej cechy, a rozmiar warstwy wyjściowej ulega zmniejszeniu. Wpływa to również znacząco na przyspieszenie procesu uczenia ze względu na redukcję parametrów.

Część klasyfikującą CNN zasadniczo można określić jako klasyczną sieć neuronową zbudowaną z kilku warstw gęstych. Pierwsza warstwa części klasyfikującej stanowi spłaszczenie danych wyjściowych warstwy poprzedzającej do jednowymiarowej tablicy. Za nią występuje kilka, w pełni połączonych, warstw ukrytych. Funkcją aktywacji stosowaną w warstwie wyjściowej CNN jest funkcja softmax. Sprawdza się ona w zagadnieniach klasyfikacji ze względu na swoje właściwości [2]. Funkcja ta przeskalowuje wartości wyjściowe tak, aby suma aktywacji wszystkich neuronów na wyjściu wynosiła 1. Tym samym wartości te stanowią prawdopodobieństwo z jakim obiekt został zaklasyfikowany do poszczególnych klas.

3. Architektura modelu

Ze względu na bardzo wysoką dokładność uzyskiwanych wyników [5], do budowy modelu do rozpoznawania ras psów na podstawie zdjęcia wykorzystano architekturę InceptionV3 (rysunek 2). Sieć ta przyjmuje na wejściu kolorowy obraz o rozmiarach

299×299 pikseli. Składa się z 22 warstw przetwarzających, w tym 42 warstw głębokich. Sieć InceptionV3 zbudowana jest z modułów Incepcji, będących połączeniem operacji splotu dla filtrów o różnych rozmiarach. Wyniki wszystkich operacji splotu są łączone na wyjściu danej warstwy.



Rysunek 2. Architektura sieci InceptionV3 [6]

W sieci InceptionV3 występują trzy rodzaje modułów Incepcji – moduł A, B oraz C. Moduły te różnią się budową i zastosowanymi filtrami. Moduły A i B pogłębiają architekturę, podczas gdy moduł C poszerza ją. W pierwszej wersji sieci Inception (InceptionV1), w modułach Incepcji, do operacji splotowych stosowane były filtry: 1×1, 3×3 i 5×5. W sieci InceptionV3, zastosowano faktoryzację, która miała na celu redukcję wymiarowości sieci. W wyniku faktoryzacji filtr 5×5 został zastąpiony dwoma filtrami 3×3. W głębszych warstwach (moduły B), dobre rezultaty przyniosło również zastąpienie filtrów n×n (n = 7) splotami 1×n oraz n×1 [7].

Na potrzeby realizowanego zagadnienia, architektura sieci InceptionV3 została zmodyfikowana. Wejście zmieniono tak, aby przyjmowało kolorowe zdjęcie o rozmiarach 224×224 pikseli. Górne warstwy sieci z części klasyfikującej zostały zastąpione warstwami:

- Global Average Pooling 2D do redukcji parametrów modelu,
- warstwą gęstą o 512 neuronach i funkcji aktywacji ReLU,

- warstwą wyjściową o 120 neuronach i funkcji aktywacji Softmax.

Tak zbudowany model składał się z 22 913 432 parametrów, z czego 1 110 648 stanowiły parametry części klasyfikującej.

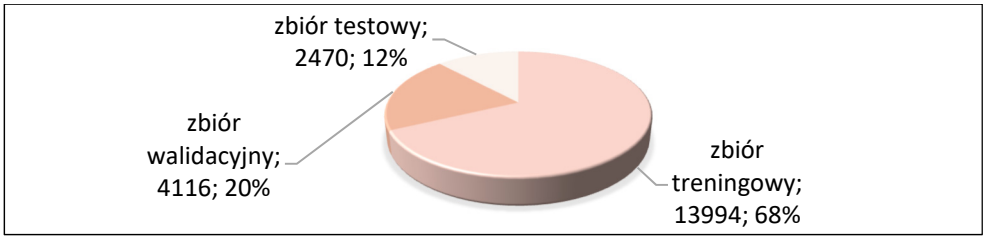
Model został napisany w języku Python 3.7. Do jego budowy wykorzystano biblioteki z zakresu uczenia maszynowego (TensorFlow 1.14.0, Keras 2.3.1, Scikit-learn 0.22.1) oraz biblioteki do obliczeń z zakresu algebry liniowej (NumPy) i wizualizacji danych (Matplotlib).

4. Proces uczenia

4.1. Dane

Do budowy modelu rozpoznającego rasy psów wykorzystano zbiór danych udostępniony przez Uniwersytet Stanforda [8] [9] [10]. Zbiór zawiera zdjęcia 120 ras psów. Liczba zdjęć dla poszczególnych ras oraz ich rozdzielczość jest różna. Najmniej licznym jest zbiór zdjęć rasy redbone – 148 zdjęć, a najbardziej licznym, zbiór zdjęć rasy maltańczyk – 252 zdjęcia. Łącznie zbiór składa się z 20 580 zdjęć wszystkich ras.

Zdjęcia zostały podzielone na trzy zbiory: treningowy, walidacyjny i testowy (rysunek 3). Zgodnie z zaleceniami, do zbioru walidacyjnego zostało zaliczonych 20% losowo wybranych zdjęć [11]. Pozostałe zdjęcia zostały przydzielone do zbioru treningowego, spośród których został wydzielony dodatkowy zbiór testowy. Liczba zdjęć tego zbioru stanowiła 15% przypadków zbioru treningowego, czyli ok. 12% wszystkich przypadków uczących.



Rysunek 3. Podział danych na zbiór treningowy, walidacyjny i testowy

Przed rozpoczęciem procesu uczenia sieci, zdjęcia ze wszystkich trzech zbiorów zostały przycięte do rozmiarów prostokątów ograniczających psy na obrazach oraz dopasowane do wejścia sieci (przeskalowane do rozmiaru 224×224 pix). Zdjęcia dostarczane na wejście sieci to obiekty o wymiarach 224×224×3. Każdy piksel obrazu został przedstawiony jako 3-elementowy wektor, którego elementy odpowiadały poszczególnym wartościom składowych modelu RGB. Dodatkowo zdjęcia zostały znormalizowane, tak aby elementy odpowiadające pikselom przyjmowały wartości z zakresu [0, 1]

Etykiety zdjęć również podlegały konwersji do postaci numerycznej. Lista nazw ras została uporządkowana alfabetycznie i każda z ras otrzymała swój odpowiednik numeryczny w postaci liczby z zakresu [0, 119]. Następnie etykiety zdjęć ze zbioru wszystkich przypadków uczących zostały zastąpione 120-elementowym wektorem, w którym element zgodny z numerem rasy (numerując od 0), otrzymał wartość 1, a pozostałe elementy wartość 0.

W celu uniknięcia zjawiska przeuczenia sieci i tym samym zwiększenia możliwości generalizacji budowanego modelu, na zbiorze danych treningowych została wykonana augmentacja danych (ang. *on-the-fly data augmentation*). Podczas procesu uczenia, na wejście sieci zamiast oryginalnych zdjęć ze zbioru treningowego, dostarczane były zdjęcia losowo zmodyfikowane (rysunek 4).

Rysunek 4. Augmentacja danych – po lewej zdjęcie oryginalne, po prawej – zdjęcia zmodyfikowane[1]

[1] Zdjęcie psa wykorzystane na rysunku (z lewej strony) zaczerpnięte ze zbioru danych udostępnionych przez Uniwersytet Stanforda na stronie: http://vision.stanford.edu/aditya86/Image-NetDogs/, listopad 2018 [8] [9] [10]. Modyfikacje (cztery modyfikacje z prawej strony) wykonano we własnym zakresie.

Modyfikacja zdjęć była wykonywana w oparciu o parametry:

- obrót zdjęcia o kąt z zakresu (-30^0, $+30^0$),
- translacja obrazu w poziomie o 0,2 szerokości obrazu,
- translacja obrazu w pionie o 0,2 wysokości obrazu,
- zmiana jasności obrazu w przedziale (0,3, 0,95),
- pochylenie obrazu o 15^0,
- powiększenie obrazu (zoom) w przedziale (0,8, 1,2),
- odwrócenie obrazu w poziomie,
- wypełnienie „pustych" pikseli, powstałych na skutek modyfikacji obrazu, wartością najbliżej położonego piksela.

Ocena sieci po procesie uczenia została przeprowadzona na niezmodyfikowanych danych testowych.

4.2. Trening sieci

Ze względu na charakter analizowanego zagadnienia oraz dużą liczbę ras, do których można zaklasyfikować psa przedstawionego na obrazie, jako funkcję straty (funkcję celu) zastosowano kategoryzacyjną entropię krzyżową. Jest to funkcja opisana wzorem:

$$-\sum_{c=1}^{M} y_{i,c} log\left(p_{i,c}\right), \tag{1}$$

gdzie:

M oznacza liczbę klas,

$y_{i,c}$ określa czy c jest prawidłową klasą dla obserwacji i przyjmując wartości 0 lub 1,

$p_{i,c}$ jest prawdopodobieństwem obserwacji i dla klasy c.

Optymalizacja kategoryzacyjnej entropii krzyżowej ma na celu minimalizację odległości między przewidywanym rozkładem prawdopodobieństwa (dostarczonym przez sieć), a rozkładem rzeczywistym. Optymalizatorem użytym do modyfikowania sieci był algorytm Adam (ang. *Adaptive Moment Estimation*) o współczynniku uczenia równym 0.001. Algorytm ten jest odmianą metody stochastycznego spadku wzdłuż gradientu (ang. *Stochastic Gradient Descent, SGD*), w której to aktualizacja

wag następuje po podaniu na wejście sieci określonej liczby przypadków zbioru treningowego (ang. *batch size*), zamiast na zakończenie epoki obejmującej cały zestaw danych treningowych. W związku z tym, gradient ma początkowo charakter przybliżony (stochastyczny) i dopiero po znacznej liczbie iteracji będzie zbliżony lub nawet zbiegnie się z rzeczywistym gradientem w minimum funkcji (lokalnym lub globalnym). Liczba próbek w niniejszym modelu została określona jako 64 przypadki treningowe. W odróżnieniu od innych algorytmów optymalizujących funkcję kosztu, algorytm Adam modyfikuje wagi nie tylko w oparciu o aktualny gradient, ale również i poprzednie gradienty.

Dodatkowo podczas budowy modelu została zdefiniowana redukcja współczynnika uczenia. Zmianę współczynnika opisano wzorem:

$$nowy\ współczynnik\ uczenia = 0{,}2 \cdot bieżący\ współczynnik\ uczenia \quad (2)$$

Redukcja współczynnika uczenia następowała w przypadku, gdy przez 4 kolejne epoki funkcja straty na zbiorze walidacyjnym nie osiągała lepszych wartości w stosunku do poprzednich epok. Do modelu został również dodany dropout – wskaźnik zapobiegający przeuczeniu sieci. W trakcie trenowania sieci, 30% losowo wybranych neuronów w danej iteracji nie podlegało uczeniu.

Szkolenie sieci zostało przeprowadzone na serwerze z systemem operacyjnym Windows Server 2016, procesorem Intel(R) Xeon(R) CPU E5-2630 v3 (2,40 GHz) oraz pamięcią RAM: 256 GB. Biorąc pod uwagę fakt, iż serwer nie posiada kart graficznych oraz fakt, że zbiór uczący nie jest licznym zbiorem, w ramach uczenia został zastosowany transfer wiedzy (ang. *transfer learning*). Technika ta zakłada, że jeśli sieć była trenowana na bardzo dużym zbiorze danych, to może ona zostać wykorzystana w innym, najczęściej podobnym, zagadnieniu. Wagi takiej sieci stanowią punkt wyjściowy procesu uczenia w kolejnych zadaniach i w zależności od wielkości zbiorów uczących, wszystkie lub tylko część z nich może podlegać modyfikacjom. Sieć InceptionV3 była trenowana na zbiorze ImageNet liczącym 14 197 122 zdjęć z 27 różnych kategorii, w tym 3822 podkategorii zwierząt, z których każda liczy średnio 732 zdjęcia.

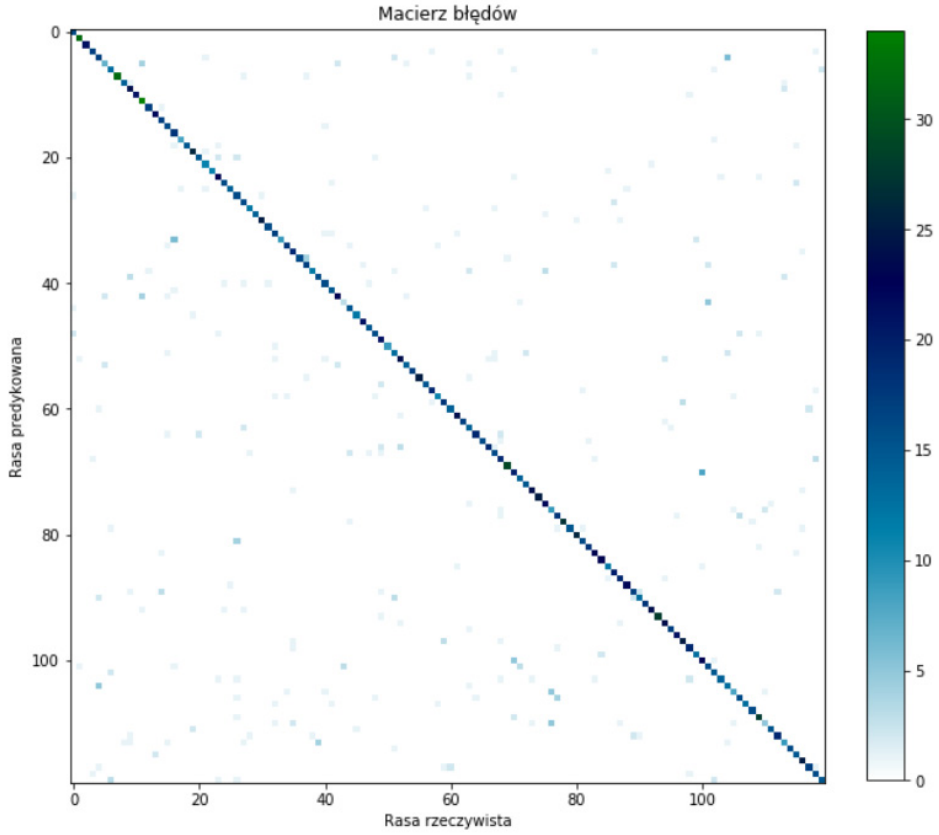
W związku z powyższym, wagi sieci InceptionV3 zostały wykorzystane do realizacji zadania rozpoznawania ras psów. Ze względu na stosunkowo nieduży zbiór uczący, uczeniu podlegały wyłącznie warstwy części klasyfikującej, a samo

trenowanie sieci zostało ograniczone do 20 epok. W trakcie trenowania sieci współczynnik uczenia był zmieniany dwukrotnie – po 7 i 19 epoce, wpływając korzystnie na osiągane później wartości metryk zarówno na zbiorze treningowym jak i walidacyjnym. W efekcie, po zakończonym procesie uczenia, dokładność modelu na zbiorze walidacyjnym osiągnęła wartość 83,94%, a funkcja straty 0,6209.

4.3. Testy modelu

Po zakończeniu uczenia sieci zostały przeprowadzone testy modelu na uprzednio wydzielonym zbiorze testowym, zawierającym zdjęcia, których sieć nie widziała w trakcie uczenia. W przypadku tego zbioru, dla obu metryk model osiągnął nieco lepsze wyniki niż w przypadku zbioru walidacyjnego. Dokładność jaką osiągnął model wyniosła 85,42%, a funkcja straty: 0,5880. Otrzymane wyniki potwierdzają, że sieć nie uległa zjawisku przeuczenia. W ramach analizy, które rasy w trakcie testów były rozpoznawane przez sieć prawidłowo, a co do których sieć popełniała błędy, została sporządzona macierz błędów (rysunek 5).

Analiza macierzy pomyłek wykazała, że sieć w trakcie testów rozpoznała bezbłędnie 20 ras psów, m.in. komondor, nagi pies meksykański, chart perski, szpic wilczy, chihuahua, likaon pstry, cyjon rudy. Wszystkie te rasy są rasami charakterystycznymi, które nie sposób pomylić z jakąkolwiek inną rasą. Najgorzej rozpoznawaną przez sieć rasą w trakcie testów był pudel miniaturowy – 47% prawidłowych predykcji. Była to jedyna rasa, dla której odsetek prawidłowych predykcji wynosił poniżej 50%. Warto zaznaczyć, że rasa ta była mylona przez sieć z rasami: pudel duży (26,5% predykcji) oraz pudel toy (26,5% predykcji). Cechy charakterystyczne pudla miniaturowego są właściwe takie same jak pudla dużego czy pudla toy. Jedyną różnicą jest wzrost jaki osiągają przedstawiciele tych ras. Zdjęcia, na których prowadzony był trening sieci, zostały wcześniej przycięte do rozmiaru prostokąta ograniczającego psa, przez co zostały pozbawione punktów odniesienia, na podstawie których możliwe byłoby oszacowanie rozmiaru psa. Wskazanie prawidłowej rasy na takich zdjęciach byłoby problematyczne nawet dla człowieka. W związku z powyższym znajduje uzasadnienie tak niski odsetek prawidłowych predykcji rasy pudel miniaturowy.



Rysunek 5. Macierz błędów[2] [3]

Zbiory zdjęć zarówno treningowych jak i testowych poszczególnych ras nie były zbiorami równolicznymi. Rasą, dla której zbiór wszystkich zdjęć (treningowych, walidacyjnych i testowych) był najmniej liczny, była rasa redbone coonhound – 148 zdjęć. Z kolei rasą o najbardziej licznym zbiorze była rasa maltańczyk – 252 zdjęć.

Analiza macierzy pomyłek wykazała, że liczba zdjęć poszczególnych ras wykorzystanych w procesie uczenia nie miała znaczącego wpływu na liczbę prawidłowych predykcji w procesie testowania modelu. W procesie uczenia sieci zostało wykorzystanych 127 zdjęć rasy redbone coonhound oraz 226 zdjęć rasy maltańczyk. Pomimo dość dużej różnicy zdjęć, odsetek prawidłowych predykcji w fazie testów

[2] Celem zapewnienia czytelności macierzy, wiersze i kolumny zostały opisane numerycznymi odpowiednikami ras psów. Zestawienie ras i odpowiadających im etykiet numerycznych zawarto w [3]. Skala barwna przedstawia liczbę predykcji.

był wysoki dla obu ras – redbone coonhound: 80,95%, maltańczyk: 92,31%. W przypadku rasy pudel miniaturowy, którą model rozpoznawał najsłabiej, podczas trenowania sieci zostało wykorzystanych 136 zdjęć. Dokładnie tyle samo zdjęć zostało użytych dla rasy borzoi, którą model w trakcie testów rozpoznawał bezbłędnie. W przypadku czterech innych ras, które model również rozpoznawał w 100% poprawnie, liczba zdjęć użyta w trakcie uczenia sieci była jeszcze niższa (np. nagi pies meksykański – 130 zdjęć).

Tabela 1. Analiza predykcji wykonanych na zbiorze testowym [3]

Cecha	Min	Max	Średnia	Mediana
Liczba predykcji	6	43	21	20
Liczba prawidłowych predykcji	3	34	18	17
Odsetek prawidłowych predykcji	47%	100%	86%	88%
Liczba ras z jaką mylona była prawidłowa rasa	0	8	2	2

Średnio w trakcie testów było wykonywanych 21 predykcji dla danej rasy, z czego średnio aż 18 było poprawnych (tabela 1). Analiza macierzy błędów pozwoliła ustalić, że średnio rasa była mylona z dwiema innymi rasami. W przypadku rasy lakeland terrier, wśród błędnych predykcji znalazło się aż 8 różnych ras, z czego połowa to inne rasy psów typu terier. Większość tych błędnych wskazań to pojedyncze pomyłki. Jedynie rasa kerry blue terrier została wskazana błędnie dwa razy. W związku z tym, nie da się rzetelnie ocenić czy sieć nie nauczyła się wystarczająco dobrze rozpoznawać rasę lakeland terrier. Możliwe, że nie wszystkie typowe cechy były widoczne na danym zdjęciu i sieć pomyliła rasy ze względu na charakterystyczny kolor umaszczenia. W przypadku analizy zdjęć pozostałych ras, większość błędnych predykcji również stanowiły jednorazowe pomyłki. Oznacza to, że sieć pomyliła daną rasę z inną wybraną rasą tylko jeden raz. Takie jednorazowe błędne wskazania stanowiły 71% wszystkich błędnych predykcji i mogą wynikać ze słabej jakości zdjęcia, niemniej jednak w celu wyciągnięcia bardziej precyzyjnych wniosków, niezbędna jest analiza jakościowa konkretnych zdjęć, dla których nastąpiła błędna predykcja.

5. Badanie wpływu jakości zdjęcia na wynik predykcji

5.1. Modyfikacja zdjęć

W ramach dodatkowych testów modelu sprawdzono również jak jakość zdjęcia podawanego na wejście wytrenowanego modelu wpływa na wynik predykcji. Weryfikacji podlegały następujące zagadnienia:

- przycięcie zdjęcia,
- rotacja zdjęcia,
- rozmycie zdjęcia,
- nałożenie efektów artystycznych na zdjęcie.

W pierwszym teście sprawdzono, jak przycięcie zdjęcia wpływa na wynik predykcji. Analizie podlegało zdjęcie psa rasy beagle, który znajdował się w centralnej części zdjęcia i stanowił niewielki element całości obrazu. W wyniku predykcji za najbardziej prawdopodobną została uznana niewłaściwa rasa toy terrier (tabela 2).

Prawidłowa rasa została wskazana jako szósta z możliwych ze znikomym prawdopodobieństwem – zaledwie 0,50%. Z analizy wyników predykcji widać, że rozpoznanie rasy było dla sieci problematycznym zadaniem. Stopień pewności, z jakim sieć wskazała najbardziej prawdopodobną rasę, jest niewielki – 38,48%.

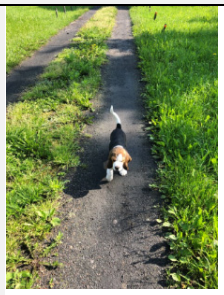
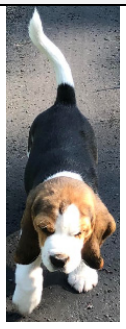
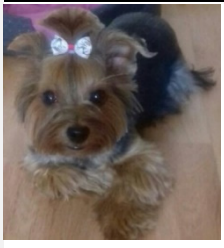
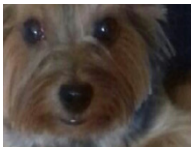
Analizując pięć ras o największym stopniu prawdopodobieństwa można przypuszczać, że sieć rozpoznała dwie cechy charakterystyczne – umaszczenie oraz niewielki wzrost. Każda z tych ras charakteryzuje się bardzo podobnym umaszczeniem stanowiącym połączenie białego, czarnego oraz brązowego koloru. Dodatkowo każda z tych ras, oprócz berneńskiego psa pasterskiego, cechuje niewielki rozmiar. W przypadku tej jednej rasy sieć mogła uznać, że na zdjęciu przedstawione jest szczenię. Przycięcie obrazu do rozmiaru prostokąta ograniczającego psa znacząco wpłynęło na wynik predykcji. Model wskazał prawidłową rasę z bardzo wysokim prawdopodobieństwem – 92,47%.

Pozostałe cztery rasy również łączy wspólna cech, ‚jaką jest podobne umaszczenie, niemniej jednak oszacowane przez sieć prawdopodobieństwo jest znikome. W tym przypadku przycięcie zdjęcia miało pozytywny wpływ na wynik predykcji, jednak kolejny eksperyment wykazał, że zbyt duże przycięcie zdjęcia może spowodować utratę cech charakterystycznych rasy, niezbędnych do uzyskania prawidłowego wyniku predykcji. Analiza oryginalnego zdjęcia psa rasy yorkshire terrier dostarczyła prawidłowego

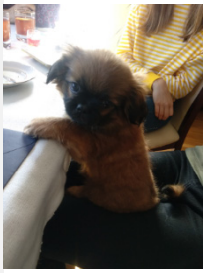
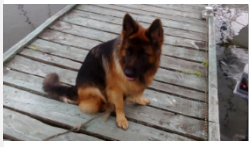
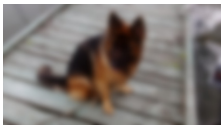
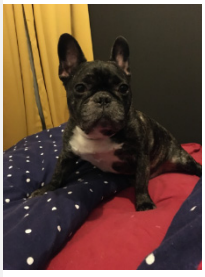
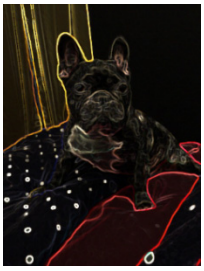
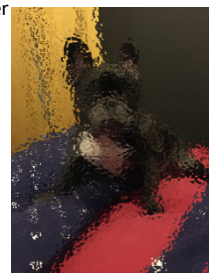
wyniku predykcji z prawdopodobieństwem graniczącym niemal z pewnością – 97,03%. Po nadmiernym przycięciu zdjęcia, w ramach którego zostały utracone informacje związane z sylwetką psa, umaszczeniem czy uszami, sieć zaklasyfikowała psa do błędnej rasy z bardzo dużym prawdopodobieństwem – 91,80%. Uwagę zwraca fakt, że prawidłowa rasa została wskazana jako druga z możliwych.

Kolejny eksperyment weryfikował wpływ rotacji zdjęcia na wynik predykcji. W wyniku analizy oryginalnego zdjęcia przedstawiającego pekińczyka, sieć rozpoznała prawidłową rasę z niemal całkowitą pewnością – 99,77%. Sieć wskazała prawidłową rasę, pomimo że pies przedstawiony na zdjęciu znajduje się w nienaturalnej pozycji.

Tabela 2. Wpływ modyfikacji obrazu na wynik predykcji[3]

Test	Predykcja			
	Zdjęcie oryginalne		Zdjęcie zmodyfikowane	
Przycięcie obrazu	 **beagle**	Predykcja: 1. 38,48% toy terrier 2. 24,58% welsh corgi cardigan 3. 17,16% berneński pies pasterski 4. 11,51% papillon 5. 5,92% welsh corgi pembroke 6. **0,50% beagle**		Predykcja: 1. **92,47% beagle** 2. 3,93% bernardyn 3. 2,21% welsh corgi cardigan 4. 0,65% treeing Walker coonhound 5. 0,46% foxhound angielski
	 **yorkshire terrier**	Predykcja: 1. **97,03% yorkshire terrier** 2. 2,16% australijski silky terier 3. 0,74% terier australijski 4. 0,04% norfolk terier 5. 0,03% toy terrier		Predykcja: 1. 91,80% terrier australijski 2. **5,43% yorkshire terrier** 3. 2,36% australijski silky terrier 4. 0,22% norwich terier 5. 0,10% norfolk terrier

[3] Zestawienie wyników predykcji zostało ograniczone do pięciu najbardziej prawdopodobnych ras.

Test	Predykcja			
	Zdjęcie oryginalne		**Zdjęcie zmodyfikowane**	
Rotacja obrazu	 **pekińczyk**	Predykcja: **1. 99,77% pekińczyk** 2. 0,23% gryfonik brukselski 3. 0,00% szpic miniaturowy 4. 0,00% shih-tzu 5. 0,00% leonberger		Predykcja: 1. 97,77% szpic miniaturowy **2. 2,13% pekińczyk** 3. 0,08% chow 4. 0,01% szpic wilczy 5. 0,00% schipperke
				Predykcja: **1. 100,00% pekińczyk** 2. 0,00% gryfonik brukselski 3. 0,00% mops 4. 0,00% shih-tzu 5. 0,00% leonberger
Rozmycie obrazu	 **owczarek niemiecki**	Predykcja: **1. 100,00% owczarek niemiecki** 2. 0,00% kelpie 3. 0,00% malinois 4. 0,00% airedale terier 5. 0,00% pinczer miniaturowy		Predykcja: 1. 54,19% toy terrier **2. 39,67% owczarek niemiecki** 3. 1,81% welsh corgi cardigan 4. 1,29% kelpie 5. 1,20% pinczer miniaturowy
Nałożenie efektu artystycznego	 **buldog francuski**	Predykcja: **1. 100,00% bulldog francuski** 2. 0,00% boston terrier 3. 0,00% chihuahua 4. 0,00% mops 5. 0,00% gryfonik brukselski		Predykcja: **1. 90,15% bulldog francuski** 2. 1,67% sznaucer miniaturowy 3. 1,31% bokser 4. 1,18% boston terier 5. 0,81% papillon
				Predykcja: 1. 23,57% chihuahua 2. 20,84% owczarek niemiecki 3. 18,96% pinczer miniaturowy 4. 4,56% kelpie 5. 4,19% papillon … **42. 0,15% bulldog francuski**

Po obrocie zdjęcia o kąt 30^0, wyniki predykcji poprawiły się. Prawidłowa rasa została wskazana ze stopniem pewności 100%. Na uzyskanie lepszego wyniku może mieć wpływ fakt, że w skutek obrotu zdjęcia, głowa psa, która dostarcza najwięcej informacji o cechach charakterystycznych, znalazła się w pozycji naturalnej.

Obrót zdjęcia o kąt 90^0 względem oryginalnego, znacząco pogorszył wyniki predykcji. Sieć wskazała błędną rasę z bardzo dużym prawdopodobieństwem – 97,77%. Prawidłowa rasa znalazła się na drugim miejscu wśród pięciu najbardziej prawdopodobnych ras. Warto również zaznaczyć, że podczas treningu sieci, wskutek augmentacji danych dozwolony był obrót zdjęcia o kąt z zakresu (-30^0, $+30^0$). Możliwe jest zatem, że tak zdefiniowany zakres miał wpływ na wynik predykcji zdjęcia obróconego o 90^0 względem oryginału.

W ramach następnego testu został zbadany wpływ rozmycia zdjęcia na wynik predykcji. Na oryginalnym zdjęciu przedstawiającym owczarka niemieckiego, sieć rozpoznała prawidłową rasę ze stopniem pewności równym 100%. Wskutek rozmycia zdjęcia krawędzie stały się niewyraźne, przez co sieć miała trudność z identyfikacją cech charakterystycznych. W wyniku predykcji została wskazana błędna rasa (toy terrier) z prawdopodobieństwem 54,19%, a prawidłowa rasa znalazła się na drugim miejscu wśród pięciu najbardziej prawdopodobnych ras. Należy zwrócić uwagę na fakt, że wszystkie te rasy mają cechę wspólną jaką jest charakterystyczne umaszczenie (domieszki czarnego i brązowego koloru). Najprawdopodobniej cecha ta była kluczową cechą, na której została oparta predykcja.

Ostatni test sprawdzał jak sieć radzi sobie z predykcją w przypadku nałożenia na zdjęcie efektów artystycznych. W ramach analizy niezmodyfikowanego zdjęcia przedstawiającego buldoga francuskiego, sieć rozpoznała prawidłową rasę ze stuprocentową pewnością. W przypadku nałożenia na zdjęcie efektów artystycznych uwypuklających krawędzie, sieć rozpoznała prawidłową rasę z bardzo dużym prawdopodobieństwem (90,15%), pomimo nałożenia na zdjęcie nienaturalnych kolorów. W drugim przypadku, na zdjęcie została nałożona tekstura, wskutek czego nastąpiło rozmycie krawędzi. Modyfikacja ta spowodowała, że wynik predykcji był błędny.

Za najbardziej prawdopodobną została uznana rasa chihuahua. Warto zauważyć, że sieć miała duże problemy ze zidentyfikowaniem rasy, o czym może świadczyć stopień pewności z jakim została rozpoznana rasa uznana za najbardziej prawdopo-

dobną – zaledwie 23,57%. Prawidłowa rasa została wskazana na 42 miejscu ze znikomym stopniem pewności – 0,15%. Analizując 5 ras uznanych za najbardziej prawdopodobne można stwierdzić, że sieć zidentyfikowała charakterystyczne duże uszy psa na zdjęciu i uwzględniła tę cechę w ramach predykcji. Rozmycie pyszczka utrudniło wydobycie innych przydatnych do analizy cech charakterystycznych.

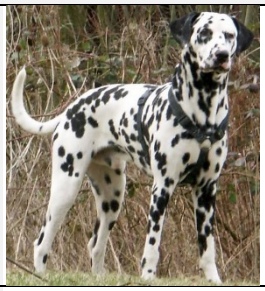
5.2. Obrazy nietypowe

Oprócz badania wpływu modyfikacji zdjęć na wyniki predykcji, zostały również zbadane zdolności predykcji modelu, w przypadku gdy na wejście dostarczane są obrazy nietypowe. W ramach jednego z testów analizie podlegało zdjęcie dalmatyńczyka. Rasa ta nie była uwzględniona w zbiorze przypadków uczących. W związku z powyższym sieć musiała wskazać błędną rasę. Sieć rozpoznała na zdjęciu rasę seter angielski z bardzo wysokim prawdopodobieństwem – 92,83% (tabela 3). Należy zwrócić uwagę na fakt, że jednym z możliwych umaszczeń seterów angielskich jest cętkowane, czarno-białe umaszczenie. W związku z tym, można stwierdzić, że rozpoznana przez sieć rasa nie została dopasowana w sposób przypadkowy, lecz wskutek identyfikacji cechy charakterystycznej.

Kolejnym obrazem wykorzystanym do testów modelu było zdjęcie zabawki psa rasy berneński pies pasterski. Zabawka przedstawiona na zdjęciu wiernie oddaje wszystkie cechy charakterystyczne psa tej rasy. W związku z tym, sieć nie miała najmniejszych problemów ze wskazaniem prawidłowej rasy. Zabawka została sklasyfikowana jako berneński pies pasterski ze stopniem pewności 100%.

Ostatnim zdjęciem wykorzystanym do testów modelu był obraz przedstawiający rysunek psa rasy yorkshire terrier. Rysunek ten został bardzo starannie wykonany i tak jak w przypadku zdjęcia zabawki, wiernie odzwierciedla wszystkie cech charakterystyczne tej rasy. W wyniku predykcji została wskazana prawidłowa rasa z bardzo wysokim prawdopodobieństwem – 94,06%. Pozostałe rasy jakie wskazał model w ramach predykcji to australijski silky terier (3,21%) oraz terier australijski (2,73%). Są to te same rasy, które wskazywał model podczas analizy zdjęcia psa rasy yorkshire terrier (tabela 3). Świadczy to o tym, że zdjęcie jest dobrej jakości oraz że sieć pomyślnie przeszła proces uczenia i umie rozpoznawać cechy charakterystyczne ras psów.

Tabela 3. Wyniki predykcji dla obrazów nietypowych[4]

Obrazy nietypowe			
Test	**Rasa spoza zbioru przypadków uczących**	**Zabawka**	**Rysunek**
Obraz			
Rasa	dalmatyńczyk	berneński pies pasterski	yorkshire terrier
Predykcja	seter angielski: 92,83%	berneński pies pasterski: 100,00%	yorkshire terrier: 94,06%

Wszystkie przeprowadzone testy udowodniły, że istotną rolę w procesie predykcji pełnią krawędzie. Im bardziej są one wyraźne, tym lepiej sieć rozpoznaje rasę psa. Wyniki predykcji wykonanych w ramach testów potwierdziły również, że proces uczenia sieci zakończył się pozytywnie i o ile zdjęcia są dobrej jakości, sieć prawidłowo klasyfikuje rasy psów prezentowanych na obrazach.

6. Podsumowanie

Głębokie, splotowe sieci neuronowe bardzo dobrze sprawdzają się w zagadnieniach analizy obrazu. Ich specyficzna architektura znacząco ogranicza liczbę parametrów sieci, przez co proces uczenia przebiega dużo szybciej. Zastosowanie transferu wiedzy w oparciu o sieć, która była trenowana na bardzo dużym zbiorze danych do realizacji podobnych zagadnień, może jeszcze dodatkowo przyśpieszyć proces uczenia. Realizacja

[4] Zdjęcie psa pochodzi ze strony: https://pl.wikipedia.org, czerwiec 2020, zdjęcie maskotki psa: https://www.pluszaki.com, czerwiec 2020, a rysunek psa pochodzi ze strony: https://malinowska.pl/galeria-portretow/, czerwiec 2020.

zadania wykazała, że wykorzystanie transferu wiedzy znacząco ułatwia budowanie sieci i wpływa bardzo korzystnie na osiągane wyniki uczenia. Zbudowana sieć osiągnęła dokładność powyżej 80% po zaledwie 20 epokach, a sam proces uczenia nie trwał długo biorąc pod uwagę fakt, iż trening sieci został przeprowadzony na serwerze nieposiadającym kart graficznych. Dodatkowym czynnikiem wpływającym na bardzo dobre rezultaty uczenia sieci były z pewnością zabiegi zapobiegające przeuczeniu sieci, takie jak augmentacja danych.

Przeprowadzone testy wytrenowanego modelu potwierdziły kluczową rolę jakości zdjęć na wynik predykcji. Model osiągał dość dobre wyniki klasyfikacji zarówno dla typowych, jak i nietypowych zdjęć. Eksperymenty wykazały również, że ekstrakcja cech charakterystycznych bazuje na wyszukiwaniu krawędzi. Dodatkowo niejednokrotnie wyniki testów dowiodły, że w realizacji niniejszego zadania, polegającego na rozpoznawaniu ras psów, istotną rolę odgrywał także kolor. W związku z tym uczenie sieci dla tego rodzaju zagadnień powinno odbywać się na zdjęciach o liczbie kanałów równej 3 (RGB).

Literatura

[1] R. Tadeusiewicz, *Sieci neuronowe*, wydanie drugie red., Warszawa: Akademicka Oficyna Wydawnicza, 1993, 8-25.

[2] K.J. Piczak, *Klasyfikacja dźwięku za pomocą splotowych sieci neuronowych*, Warszawa: Politechnika Warszawska, 2018, 15-75.

[3] A. Sierżantowicz, *Opracowanie koncepcji i implementacja modelu rozpoznawania obrazu dla zadanej dziedziny z wykorzystaniem elementów sztucznej inteligencji*, praca inżynierska, Warszawa: Warszawska Wyższa Szkoła Informatyki, 2020.

[4] *Math Works*, The MathWorks, Inc., [Online]. https://www.mathworks.com/solutions/deep-learning/convolutional-neural-network.html. [październik 2020].

[5] *Keras*. [Online]. https://keras.io. [Data uzyskania dostępu: marzec 2020].

[6] S.-H. Tsang, *Review: Inception-v3 — 1st Runner Up (Image Classification) in ILSVRC 2015*, A Medium Corporation, 2018. [Online]. https://medium.com/@sh.tsang/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c. [czerwiec 2020].

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, *Rethinking the Inception Architecture for Computer Vision*, 2015. [Online]. https://arxiv.org/abs/1512.00567. [czerwiec 2020].

[8] A. Khosla, N. Jayadevaprakash, B. Yao i L. Fei-Fei, *Novel Dataset for Fine-Grained Image Categorization*, [w:] *First Workshop on Fine-Grained Visual Categorization (FGVC)*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li i L. Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database*, [w:] *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] *S.V. Lab*, Stanford Vision Lab, [Online]. http://vision.stanford.edu/aditya86/ImageNetDogs/. [listopad 2018].

[11] *Leksykon sieci neuronowych*, wydanie pierwsze (red.) R. Tadeusiewicz, M. Szaleniec, Wrocław: Wydawnictwo Fundacji „Projekt Nauka", 2015.

Development of the concept and implementation of an image recognition model using elements of artificial intelligence

Abstract

This article presents the concept and implementation of a model for recognizing dog breeds based on an input image. The task was performed with the use of a deep neural network model based on the InceptionV3 structure. The neural network has been trained and tested on a dataset counting more than 20,000 images of 120 dog breeds using transfer learning technique. The impact of image quality on classification results was also examined. The model obtained very good results in the analysis of both typical and unusual input images.

Keywords *– computer vision, artificial neural network, convolutional neural network, InceptionV3, transfer learning*