Summary in Polish

Analiza zawartości obrazu jest bardzo złożonym tematem, w którym dobór odpowiedniej metody ma bardzo duże znaczenie dla efektów końcowych. Ważnym aspektem jest to, aby móc wyekstrahować reprezenatywne cechy opisujące obraz lub obiekty na nim zawarte i zapisać je w formie deskryptorów liczbowych. Taka forma reprezentacji kluczowych elementów obrazu daje nam ogromne możliwości pod kątem wyszukiwania podobnych obrazów. Istotnym elementem tego problemu jest określenie tych obszarów analizowanego obrazu, które są dla nas najistotniejsze podczas realizacji tego zadania. W tej pracy przedstawiono nowe podejście do zagadnienia Content Based Image Retrieval, proponując metodę analizy głębokiej sieci neuronowej w zakresie interpretowania jej zawartości i zdefiniowania reguł, jakimi się kieruje w procesie decyzyjnym. Wytyczenie tych zasad pozwala na zbudowanie uniwersalnych deskryptorów umożliwiających zapis najistotniejszych cech opisujących zawartość obrazu. Sposób generowania tych deskryptorów również jest bardzo istotny, ponieważ może to narzucić nam pewne ograniczenia pod kątem odpowiedniej analizy. Zaproponowane rozwiązanie przedstawia uniwersalną strukturę deskryptora umożliwiającą dowolne modelowanie w zależności od tego, jakie informacje zawarte na obrazie chcemy analizować. Badania przedstawione w tej pracy pokazują skuteczność tego rozwiązania oraz możliwości jego wykorzystania w zadaniach Content Based Image Retrieval.

Najnowsze metody głębokiego uczenia w ostatnich latach zrobiły ogromny postęp zarówno w zastosowaniach biznesowych, jak i też w pracach badawczo-rozwojowych. Wiele struktur, które są wykorzystywane obecnie w procesie głębokiego uczenia powstały już kilkanaście lat temu, jednak dopiero współczesne możliwości obliczeniowe kart graficznych pozwalają na wykorzystanie ich rzeczywistych możliwości. W międzyczasie powstało wiele dodatkowych metod wspomagających te struktury zarówno w procesie uczenia, jak i też podczas analizy zawartości danych, przez co skuteczność tych rozwiązań w

wielu zastosowaniach okazuje się bezkonkurencyjna.

Chociaż informacje zawarte w aktywacjach warstw konwolucyjnych są bardzo przydatne w zadaniach takich jak odszumianie, segmentacja i klasyfikacja obrazów, są one jednocześnie trudne do interpretacji. Duża liczba połączeń w głębokich sieciach neuronowych oraz złożone zależności pomiędzy aktywacjami neuronów utrudniają wyodrębnienie tylko tych cech, które są najważniejsze dla danego obrazu. Niestety istniejące rozwiązania nie zapewniają idealnej ekstrakcji cech. W tej pracy opracowano nowy typ deskryptorów, rozszerzając ideę kodów neuronowych przedstawioną w [9] poprzez konkatenację najbardziej znaczących aktywacji z warstw splotowych z aktywacjami z warstw w pełni połączonych. Takie deskryptory oddzielają nieistotny szum od wartościowej wiedzy opisującej obraz. Należy zauważyć, że podobne podejście zastosowano wcześniej do generowania skrótów, czyli deskryptorów binarnych [25], [26]. Kody binarne mogą być bardzo przydatne przy pobieraniu obrazów z dużych baz danych, jednak nie mogą zagwarantować tak wysokiej dokładności jak deskryptory o wartościach rzeczywistych. Dlatego w tej pracy skupiono się na deskryptorach, których wartości są pobierane bezpośrednio z aktywacji neuronów.

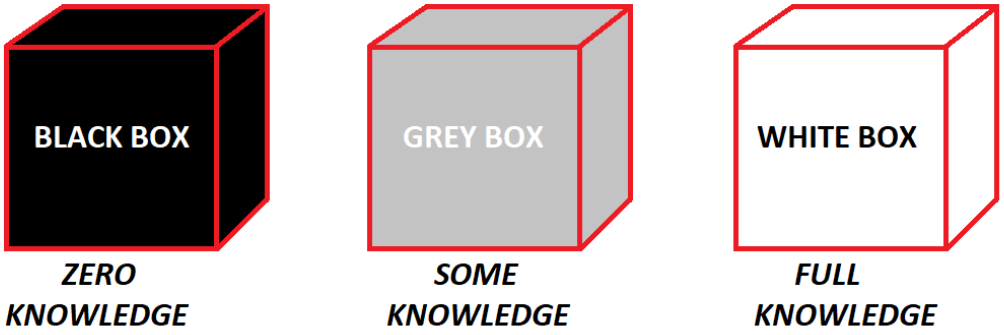
W pracy przedstawiono nasz nowatorski algorytm, który analizuje bloki warstw splotowych sieci neuronowej i wybiera sygnały najbardziej istotne z punktu widzenia decyzji klasyfikatora [34]. Ponieważ warstwy konwolucyjne zbudowane są w oparciu o filtry, ich działanie będzie silniejsze dla cech poprawiających teksturę obrazu. Łącząc takie aktywacje ze splotowej części sieci z wartościami zebranymi z sygnałów w pełni połączonych warstw, uzyskujemy zestaw cech skupiających się na semantycznej kategorii obiektów zawartych w obrazie oraz innych wspomnianych wcześniej cechach obrazu, takich jak tekstury lub elementy tła. Wygenerowany deskryptor będzie zawierał bardzo szczegółowe cechy opisujące całą zawartość obrazu. W niniejszej pracy wykorzystano powszechnie znaną sieć neuronową VGG16, ze względu na jej stosunkowo prostą budowę. Ta metoda generowania deskryptorów może być również bezpośrednio dostosowana do sieci o innych strukturach.

Generowanie optymalnych deskryptorów odzwierciedlających charakterystyczne cechy obrazów propagowanych przez głęboką sieć neuronową wymagają dobrego zrozumienia tych struktur. Pierwszym krokiem do poznania głębokich sieci neuronowych jest umiejętność odpowiedniej konfiguracji struktur względem danych oraz badanie aktywności poszczególnych warstw względem skuteczności. Takie podejście pomaga nam zrozumieć

przepływ danych wraz z aktywnością charakterystycznych cech, co jest nieodłącznym elementem budowania efektywnych deskryptorów.

Wyniki zaprezentowane w chapter 2 prezentują proces uczenia w odniesieniu do różnych konfiguracji głębokich sieci neuronowych. Wprowadzone zmiany w strukturach pokazują, jak reaguje sieć, na proces adaptacji struktury do złożoności danych. Wprowadzanie takich modyfikacji uczy nas jak projektować struktury i przede wszystkim, uświadamia nas, jak dany model działa w praktyce. Ciągle rozwijająca się dziedzina głębokich sieci neuronowych generuje bardzo dużo nowych metod oraz technik konfiguracji głębokich struktur dlatego warto na bieżąco je poznawać i analizować ich wpływ na proces uczenia i w konsekwencji na ich działanie docelowe.

Wprowadzanie modyfikacji w poszczególnych modelach, daje nam pewien obraz ich działania, jednak nie pozwala nam tak naprawdę zrozumieć co się dzieje wewnątrz. Głębokie sieci neuronowe przez ich ogromną strukturę są bardzo trudne do interpretacji, to spowodowało powstanie pojęć takich jak black box, grey box oraz white box Rysunek 1. Wiemy co się dzieje na wejściu, oraz na wyjściu głębokiej struktury, ale bardzo trudno zrozumieć jej działanie wewnątrz, dlatego od samego początku okryto je mianem Black box, czyli czymś całkowicie niezrozumiałym. Aktualna wiedza oparta na doświadczeniu oraz dostępne metody umożliwiają zrozumienie nieco więcej. Możemy zajrzeć do głębokiej struktury i przynajmniej częściowo zrozumieć co się w niej dzieje starając się wyłapać sygnały charakterystyczne dla konkretnych cech zawartych na obrazie, np. pewna charakterystyczna krawędź lub tekstura. Tak zinterpretowaną strukturę określamy mianem grey box. Jednak finalny wynik całego modelu opiera się nie tylko na pojedynczych cechach, ale również na konkretnych zależnościach pomiędzy nimi. Zrozumienie tego zagadnienia daje nam pełną interpretację danej struktury, którą możemy wtedy nazwać white box. Interpretacja tego elementu jest znacznie bardziej skomplikowana, na samym początku istnienia głębokich sieci neuronowych określana nawet jako niemożliwa do wykonania. Jak widać, jest to bardzo duże wyzwanie badawcze, wyniki zaprezentowane w pracy [34] otwierają nowe możliwości takiej interpretacji.



Rysunek 1: Poziomy interpretowalności głębokich struktur.

Generowanie deskryptorów reprezentujących charakterystyczne cechy zawarte na obrazie z wykorzystaniem głębokich sieci neuronowych, jak już wiemy, wymaga dużego doświadczenia w projektowaniu, oraz rozumieniu działania tego typu struktur. Interpretowalność ma tutaj bardzo duże znaczenie, gdyż znajduje ona bardzo wiele zastosowań w wielu dziedzinach naukowych oraz biznesowych, takich jak: przemysł samochodowy, branża medyczna, wirtualna rzeczywistość, gry, handel detaliczny, przemysł bezpieczeństwa, platformy mediów społecznościowych, wizualne wyszukiwarki oraz wiele innych. Każde z wymienionych zastosowań wymaga innego sposobu analizy zawartości obrazów. Możemy analizować szczegółowo zarówno obiekty znajdujące się na zdjęciu, jak i też jego aktywność, otoczenie, tekstury tła, drugoplanowe elementy sceny, itd. Niezależnie od tego, jakie zastosowanie oraz metodę wybierzemy, zawsze najważniejszą rolę odgrywa skuteczność. Jest wiele metod umożliwiających badanie zawartości obrazów. Znaczna większość tych metod opiera się na budowaniu pewnego rodzaju deskryptora, który wyodrębnia cechy najbardziej charakterystyczne dla danego obrazu propagowanego w danej chwili. W tej pracy przedstawiono mechanizm budowania deskryptora, który jest w pełni autorskim i zarazem uniwersalnym rozwiązaniem umożliwiającym jego adaptację do dowolnego modelu głębokiej sieci neuronowej.

Mając już pewne doświadczenie z zagadnieniem głębokich sieci neuronowych oraz samym tematem Content Based Image Retrieval możemy się zastanawiać, od czego zacząć. Punktem wyjściowym była analiza aktywności sieci podczas propagacji próbek, ponieważ te wartości odzwierciedlają pewną cząstkę całości decyzji danego modelu. Właśnie tutaj możemy znaleźć bardzo cenne informacje, które mogą pomóc w budowaniu efektywnych deskryptorów, a w rezultacie lepiej zrozumieć działanie tego typu struktur i je docelowo

zinterpretować. Podobnie jak w poprzednim rozdziale na początek przeanalizowano niewielką sekwencyjną strukturę przeznaczoną do klasyfikacji obrazów zbudowana z części warstw konwolucyjnych oraz klasyfikatora fully connected.

Deskryptory zbudowane na bazie klasyfikatora skupiają się jedynie na kategorii obiektów, co niekoniecznie musi być dla nas zadowalające. W tym celu powinniśmy zainteresować się wiedzą umieszczoną nieco głębiej w strukturze, czyli w warstwach konwolucyjnych, ponieważ właśnie tam jest analizowana cała zawartość obrazu. Konwolucyjne sieci neuronowe odgrywają największą rolę w analizie zawartości obrazów, ponieważ informacja przetwarzana przez klasyfikator umieszczony na końcu całej struktury jest znacznie bardziej uproszczona. Niestety warstwy konwolucyjne pomimo wysokiej skuteczności w dziedzinach takich jak odszumianie, segmentacji lub klasyfikacji obrazów, są bardzo ciężkie do interpretacji. Duża ilość połączeń w głębokich strukturach bardzo utrudnia separacje istotnych informacji umożliwiających zbudowanie deskryptora. Analizując aktywację neuronów, widzimy jedynie wielki szum. Konieczną rzeczą jest zastosowanie tutaj pewnego mechanizmu, który nam ułatwi znalezienie reguł pozwalających wyseparować jedynie te najbardziej znaczące dla nas informacje.

Bazując na metodzie opisanej w tej pracy, zbudowano deskryptor oparty o aktywności zebrane z warstwy fully connected oraz poprzez analizę aktywności feature maps. Takie podejście daje dodatkową informację o dodatkowych elementach zawartych w obrazie. Wyszukiwanie obrazów najbardziej podobnych względem query images prezentuje Rysunek 2. Kiedy przyjrzymy się dokładniej wynikom, widzimy, że osiągnęliśmy dokładnie to, czego się spodziewaliśmy. Podobne obrazy zawierają zarówno podobne obiekty, jak i też pozostałe elementy zawarte w obrazie.

Rysunek 2: Top-10 podobnych obrazów wygenerowane dla dziesięciu obrazów różnych klas.

Algorytm generujący deskryptory bazuje na głębokich sieciach neuronowych wraz z warstwami fully connected. Tak jak wspomniano wcześniej, generowanie deskryptorów w oparciu jedynie o warstwy fully connected nie jest efektywne, dlatego skupiamy się na całej strukturze wraz z warstwami konwolucyjnymi. Analiza wartości aktywacji neuronów w warstwach klasyfikatora pozwala skutecznie wyekstrachować jedynie kategorię obiektu, bo to jest jego konkretnym celem, bardziej szczegółowe elementy obrazu są analizowane we wcześniejszych warstwach konwolucyjnych. Skoro wiadomo, że zaproponowany algorytm w poprzednim rozdziale, wykazuje możliwości analizy warstw konwolucyjnych pod kątem ekstrakcji bardziej szczegółowych cech obrazu, zbadano jego działanie, weryfikując różne warianty modelowania deskryptora, bazując na tym podejściu. Cały algorytm został również ustandaryzowany w formie matematycznej. Dla dodatkowego porównania wykorzystano również model ResNet50, który jest znacznie bardziej złożony od modelu VGG16, za czym idzie również większa skuteczność klasyfikacji na zbiorze Image NET (VGG16 accuracy: TOP-1 = 64.72%, TOP-5 = 85.74%; ResNet50 accuracy: TOP-1 = 83.20%, TOP-5 = 96.50%). Takie podejście znacząco zwiększa wiarygodność przedstawionych wyników badań.

Zaproponowane deskryptory porównano z innymi rozwiązaniami, które można znaleźć w literaturze, a także z deskryptorami opartymi wyłącznie na aktywacjach warstw w

pełni połączonych lub splotowych. W większości przypadków proponowany deskryptor przewyższał pozostałe. Wydajność wszystkich porównywanych deskryptorów została zmierzona przy użyciu różnych metryk, tj. odległości L_1 między samymi deskryptorami, deskryptora L_1 między macierzami Gramma pobranych obrazów oraz obszaru nakładania się między estymatorami rozkładu kolorów wyznaczonymi przy użyciu funkcji jądrowych Parzena. Wyniki uzyskane dla trzech powyższych miar zostały zaprezentowane odpowiednio w Tablicach 1, 2 i 3.

TABLICA 1: Średnia odległość L_1 pomiędzy deskryptorami.

Query Image	**IMAGE NET1M** $IM(x)$	**FC** $h(x)$	**Conv.** $\tilde{\eta}(x)$	**FC + Conv.** $\eta(x)$	**Av. pool** $A(x)$	**Py. pool** $P(x)$	**Res Net50** $R(x)$
1	0.385	0.428	0.306	**0.236**	0.384	0.387	0.346
2	0.205	0.185	**0.173**	0.192	0.199	0.185	0.215
3	0.317	0.388	0.283	**0.189**	0.453	0.395	0.347
4	0.246	0.226	**0.184**	0.198	0.268	0.252	0.232
5	0.215	0.209	0.232	0.216	**0.165**	0.199	0.223
6	0.296	0.223	0.264	**0.192**	0.233	0.212	0.247
7	0.282	0.269	0.247	**0.210**	0.271	0.250	0.233
8	0.268	0.282	0.239	**0.221**	0.257	0.281	0.276
9	0.183	0.227	0.192	0.155	0.197	0.175	**0.136**
10	0.107	0.103	**0.056**	0.084	0.085	0.124	0.057
Avg.	0.250	0.254	0.218	**0.189**	0.251	0.246	0.231

Zaproponowana metoda generowania deskryptorów oparta na wyselekcjonowanych wartościach aktywacji dostarcza rozwiązanie, które uwzględnia cały kontekst obrazu propagowanego poprzez głęboką sieć neuronową. Deskryptory analizują wartości aktywacji zarówno z warstw konwolucyjnych, jak i warstw fully connected. Przedstawiona metoda umożliwia przeszukiwanie podobnych kontekstowo obrazów, uwzględniając przy tym wszystkie elementy zawarte na obrazie, ale również klasyfikowany przez model obiekt. Skuteczność proponowanych deskryptorów została przedstawiona z wykorzystaniem zbioru IMAGENET1M z przeznaczeniem do badań dla CBIR. Pomimo osiągnięcia bardzo obiecujących wyników, w większości przypadków lepszych od wcześniej opracowanych algorytmów w literaturze, przedstawiony w niniejszej pracy algorytm konstrukcji

Tablica 2: Średnie powierzchnie przekrywania się estymatorów Parzena dla rozkładów kolorów.

Query Image	**IMAGE NET1M** $IM(x)$	**FC** $h(x)$	**Conv.** $\tilde{\eta}(x)$	**FC + Conv.** $\eta(x)$	**Av. pool** $A(x)$	**Py. pool** $P(x)$	**Res Net50** $R(x)$
1	0.186	0.087	0.315	**0.470**	0.171	0.179	0.254
2	0.606	0.639	**0.652**	0.626	0.606	0.635	0.576
3	0.192	0.029	0.289	**0.516**	0.051	0.029	0.122
4	0.499	0.539	**0.619**	0.590	0.464	0.491	0.524
5	0.575	0.569	0.527	0.557	**0.651**	0.581	0.560
6	0.438	0.586	0.481	**0.624**	0.564	0.597	0.520
7	0.437	0.456	0.498	**0.565**	0.441	0.489	0.526
8	0.479	0.440	0.523	**0.563**	0.488	0.428	0.430
9	0.559	0.492	0.562	**0.644**	0.569	0.578	0.623
10	0.374	0.417	**0.525**	0.485	0.454	0.405	0.487
Avg.	0.434	0.425	0.499	**0.564**	0.226	0.441	0.462

Tablica 3: Średnia odległość L_1 pomiędzy macierzami Gramma.

Query Image	**IMAGE NET1M** $IM(x)$	**FC** $h(x)$	**Conv.** $\tilde{\eta}(x)$	**FC + Conv.** $\eta(x)$	**Av. pool** $A(x)$	**Py. pool** $P(x)$	**Res Net50** $R(x)$
1	1.282	0.659	1.122	0.934	0.714	**0.503**	0.983
2	1.328	1.360	1.267	1.196	**0.999**	1.428	1.068
3	1.465	0.926	1.277	**0.890**	0.935	0.955	0.933
4	1.041	0.975	1.271	0.877	**0.769**	1.020	0.790
5	1.445	0.675	0.588	0.607	**0.567**	0.806	0.738
6	2.486	2.404	**1.323**	1.393	2.362	1.461	2.791
7	1.627	**0.928**	1.056	1.065	1.259	1.899	1.041
8	1.055	1.072	0.941	**0.886**	1.091	1.168	1.028
9	1.212	1.104	1.177	**1.024**	1.143	1.162	1.057
10	1.582	1.745	1.380	**1.293**	1.577	1.469	1.408
Avg.	1.252	1.185	1.140	**1.017**	1.142	1.187	1.184

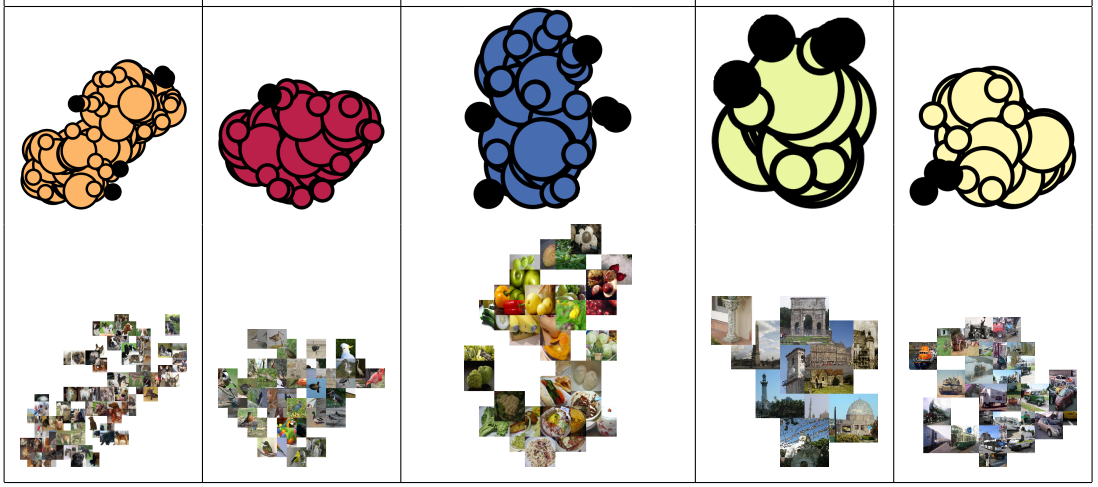
deskryptorów otwiera szeroki obszar możliwych badań z dużym potencjałem na dalszy rozwój.

Większość współczesnych techniki przetwarzania obrazów bazuje na głębokim uczeniu [47] i splotowych sieciach neuronowych (CNN). Niestety to, co się dzieje podczas procesu uczenia głębokich struktur jest zazwyczaj mało zrozumiałe. Jest to dobry powód do tego, aby spróbować przeprowadzić badania tego procesu i spróbować przeanalizować oraz zinterpretować to, co można wyodrębnić z wytrenowanej głębokiej sieci neuronowej. Metoda przedstawiona w chapter 4 prezentuje pewne możliwości interpretacji głębokich struktur poprzez zaproponowany mechanizm ekstrakcji pewnych zależności, bazujących na metodach opisanych w poprzednich rozdziałach. Metoda ta może zostać wykorzystana nie tylko dla CBIR, ale również w procesie wspomagania klasyfikacji obrazów. Reguły wygenerowanie na podstawie zaproponowanego algorytmu są pogrupowane względem różnych semantycznie powiązanych ze sobą klas, przy czym wszystkie te klasy w danej grupie należą do tej samej kategorii nadrzędnej, np. zwierzęta, pojazdy, jedzenie, budynki, itp. Ta informacja pozwala nam zwiększyć prawdopodobieństwo klasyfikacji, poprzez zawężenie obszaru poszukiwań, korzystając właśnie z tych grup.

Zaproponowaną metodę konstruowania reguł można podzielić na trzy kroki, które dokładniej zostały opisane w podrozdziale 4.1. Pierwszym krokiem jest skonstruowanie odpowiedniego deskryptora, zawierającego charakterystyczne cechy pojedynczych obrazów bazując na metodzie opisanej w chapter 3. Dla uproszczenia badań deskryptor wygenerowano na podstawie ostatniego bloku złożonego z warstwy konwolucyjnej. Drugim krokiem jest wyznaczenie średnich wartości dla deskryptorów reprezentujących poszczególne kategorie obiektów. W tym celu wykorzystano średnią arytmetyczną wszystkich deskryptorów obrazów należących do tej samej klasy. W trzecim etapie analizy zastosowano algorytm grupowania DBSCAN [48] do grupowania średnich wartości wytyczonych w drugim etapie.

Po wykonaniu wyznaczonych kroków badań zaobserwowano, że wartości poszczególnych deskryptorów grupują się w grupy powiązane ze sobą nadrzędną kategorią [49]. Grupy, które można jednoznacznie odseparować, traktowane są jako odrębne reguły wspierające CBIR, gdzie dystans pomiędzy poszczególnymi obrazami umieszczonymi w przestrzeni dwuwymiarowej reprezentuje stopień podobieństwa. Podobnie jak w poprzednich rozdziałach do symulacji wykorzystano model VGG16. Weryfikacja zaproponowanej metody

Tablica 4: Pięć przykładowych klastrów uzyskanych za pomocą algorytmu DBSCAN. Każdy klaster zawiera deskryptory obrazów różnych (ale semantycznie podobnych) klas, należących do jednej wspólnej kategorii nadrzędnej.

G1	**G2**	**G3**	**G4**	**G5**
Dogs	Birds	Fruit and Vegetables	Vehicles	Buildings
				

jest przeprowadzana na podstawie zbioru danych ILSVRC. W Tablicy 4 przedstawiono przykładowe klastry uzyskane na bazie metody opisanej powyżej.

Zaprezentowane wyniki wyglądają bardzo obiecująco i otwierają nowy obszar do dalszych badań umożliwiający wsparcie procesu klasyfikacji obrazów czy też bardziej szczegółowej interpretowalności głębokich sieci neuronowych. Wyznaczenie reguł pozwalających określić pewne korelacje pomiędzy klasami, czy też uchwycenie wartości wygenerowanych deskryptorów odzwierciedlających charakterystyczne cechy obrazu może również znacząco wspomóc zagadnienie CBIR, które jest istotną częścią tej pracy.

Podsumowując, główne nowości i cechy charakterystyczne tej rozprawy są następujące:

- Zaproponowano nowy hybrydowy algorytm konstrukcji deskryptorów, uwzględniające aktywacje neuronowe zarówno z warstw splotowych, jak też w pełni połączonych warstw głębokiej sieci neuronowej;
- Zaproponowane deskryptory umożliwiają pozyskiwanie obrazów, które są podobne do obrazu zapytania nie tylko semantycznie, ale także pod względem drugorzędnych cech obrazu, takich jak rozkład kolorów, tło, tekstury itp.;
- Zaproponowane deskryptory porównano z innymi rozwiązaniami, które można znaleźć w literaturze, a także z deskryptorami opartymi wyłącznie na aktywacjach

warstw w pełni połączonych lub splotowych. W większości przypadków proponowany deskryptor przewyższał pozostałe;

- Wydajność wszystkich porównywanych deskryptorów została zmierzona przy użyciu różnych metryk, tj. odległości L_1 między samymi deskryptorami, deskryptora L_1 między macierzami Gramma pobranych obrazów oraz obszaru nakładania się między estymatorami rozkładu kolorów wyznaczonymi przy użyciu funkcji jądrowych Parzena;

- Zagadnienie interpretowalności zastosowanych głębokich sieci neuronowych zostało zbadane za pomocą algorytmów t-SNE oraz DBSCAN. Okazało się, że deskryptory obrazu można pogrupować w klastry, które pasują do oryginalnej hierarchii klas zastosowanych zbiorów danych. Fakt ten może zostać wykorzystany w przyszłych badaniach do doskonalenia procesu klasyfikacji sieci neuronowej.

Możliwości rozwoju zagadnienia CBIR są bardzo duże i można w tej przestrzeni zdziałać jeszcze bardzo wiele. Zaprezentowane metody w tej pracy są jedynie niewielką cząstką tego, co już zostało zbadane, jednak jest to na pewno coś innego. Mam nadzieję, że ta praca będzie inspiracją do dalszych badań i pozwoli odkryć nowe nieograniczone możliwości CBIR.