Streszczenie

Prowadzone w ostatnich latach badania z zakresu sztucznej inteligencji pokazały, że systemy uczące się są kluczem do rozwiązania wielu praktycznych problemów, w szczególności w dziedzinie wizji komputerowej. Na szczególną uwagę zasługują metody uczenia głębokiego (ang. deep learning methods), których rozwój pozwolił na uzyskanie bezprecedensowych wyników w klasycznych zastosowaniach informatyki, takich jak rozpoznawanie twarzy (Schroff et al., 2015), rozpoznawanie emocji (Xia et al., 2017), wyszukiwanie obiektów (Noh et al., 2017), śledzenie obiektów (Dong and Shen, 2018), detekcja obiektów (Tan et al., 2020), generowanie zdjęć (Goodfellow et al., 2014), tłumaczenie maszynowe (Edunov et al., 2018) i innych.

Niniejsza praca jest poświęcona danym *multimodalnym*. W ogólności, *modalność* jest rozumiana jako społecznie lub kulturowo ukształtowany środek przekazu informacji, taki jak obraz, pismo, mowa lub ruchome obrazy (Kress, 2010). W związku z wykładniczym rozwojem technologii coraz więcej informacji jest przetwarzanych przez komputery. Korzystanie i rozumienie informacji multimedialnych jest jedną z funkcji wymaganych od sztucznej inteligencji. Taka umiejętność wymaga jednoczesnego korzystania z wielu modalności i wyciągania ujednoliconych wniosków z różnych typów przekazu.

Ludzie posiadają naturalne predyspozycje do łatwego manipulowania reprezentacjami multimodalnymi i rozpoznawania tych samych pojęć semantycznych, ukazanych innym sposobem przekazu. Gdy komunikujemy się z nieznajomą osobą, zauważamy jej mimikę twarzy, analizujemy ton głosu i odczytujemy mowę ciała, aby ocenić jej intencje. Zdolność ta ma kluczowe znaczenie zarówno dla podstawowych zadań takich jak komunikacja i poruszanie się w środowisku, ale również w celach twórczych i kreatywnych. Dla maszyn, na odwrót, jest to wyjątkowo skomplikowane zadanie, wymagające długiego procesu uczenia i danych pochodzących z różnych typów sygnałów.

Możemy zaobserwować rosnące zainteresowanie metodami analizy danych multimodalnych, w szczególności metodami z zakresu łączenia obrazu i tekstu. Chociaż metody głębokiego uczenia się były szeroko stosowane zarówno do danych obrazowych, jak i danych tekstowych, badania nad jednoczesnym zrozumieniem tych dwóch modalności dopiero nabierają rozpędu. Jest to spowodowane, z jednej strony, postępem i popularnością najnowszych metod w świecie nauki, takich jak DALL·E (Ramesh et al., 2021), która jest w stanie generować obrazy z zapytania tekstowego, także uwarunkowane obrazem (np. „Wygeneruj manekina ubranego w czarną skórzaną kurtkę i złotą plisowaną spódnicę"). Z drugiej strony wysoka przydatność takich rozwiązań, zwłaszcza w środowisku e-commerce, napędza zainteresowania badawcze w branży. Główne platformy handlowe, takie jak Zalando, prowadzą badania nad kodowaniem danych multimodalnych z domeny mody, a także nad generatywnym projektowaniem mody, uzależnionym od pozy ciała, określonych przedmiotów lub atrybutów (takich jak kolor lub kształt).

Innym ważnym zastosowaniem osadzania multimodalnego jest wyszukiwanie mediów. Sumaryczna ilość danych ogromnie wzrosła w ostatnich latach i nadal

rośnie. Stanowi to poważne wyzwanie dla aplikacji multimedialnych i handlu elektronicznego, które opierają się na wyszukiwaniu. Wiodące platformy internetowe, takie jak Google, Alibaba, Amazon czy Bing, zostały opracowane wokół wyszukiwarek (Zhang et al., 2018; Houdong et al., 2018). Co więcej, interaktywne wyszukiwarki multimedialne są wszechobecne w urządzeniach mobilnych i pozwalają na zapytania głosowe, tekstowe lub wizualne (Li et al., 2013; Sang et al., 2013; Chen and Girod, 2015).

Chociaż poczyniono znaczne postępy w dziedzinie wyszukiwania multimodalnego, to wciąż jest trudne zadanie ze względu na lukę semantyczną pomiędzy reprezentacjami cech (ang. *embedding*) a wysokopoziomowymi pojęciami semantycznymi. Ponadto, nietrywialnym zadaniem jest odszyfrowanie dokładnych intencji użytkownika. Korzystanie z unimodalnych wyszukiwarek (takich jak wyszukiwanie tekstem lub wyszukiwanie obrazem) ma jedną istotną wadę: uniemożliwia użytkownikowi naturalne przemieszczanie się pomiędzy modalnościami, takie jak: „Szukam tego typu sukienki jak na zdjęciu, ale uszytej z jedwabiu". Wynika to głównie z faktu, że pojęcie podobieństwa w przestrzeniach różnych modalności jest inne niż we wspólnej przestrzeni multimodalnej. Ponadto modelowanie tej wielowymiarowej przestrzeni multimodalnej wymaga bardziej złożonych strategii uczenia i bogatych zbiorów danych.

Inne rodzaje problemów, które wymagają uczenia na danych multimodalnych, to wizualne odpowiadanie na pytania (ang. visual question answering), wyszukiwanie obrazów w oparciu o tekst, automatyczne podpisywanie obrazów, rozumienie sceny, generowanie obrazu z tekstu, modyfikacja obrazu uzależniona od tekstu i inne. Warto również wspomnieć, że chociaż badania reprezentacji multimodalnych koncentrują się głównie na danych obrazowych i tekstowych, nie ograniczają się tylko do tych dwóch modalności. Prowadzone są również badania dotyczące multimodalnych reprezentacji audiowizualnych do klasyfikacji lub ulepszania mowy (Hou et al., 2018; Mroueh et al., 2015). Komunikacja człowiek-komputer ma kluczowe znaczenie dla budowy systemów autonomicznych i obejmuje szeroki zakres innych modalności, takich jak gesty ciała, sygnały biologiczne i mimika twarzy (Ranganathan et al., 2016). W rozdziale 7 pokazujemy, jak dodanie sygnału z dodatkowej modalności w wielozadaniowym scenariuszu uczenia się poprawia klasyfikację emocji ze zdjęcia twarzy.

Niniejsza praca jest poświęcona opracowaniu lepszych reprezentacji multimodalnych, które można wykorzystać do szeregu zadań, takich jak klasyfikacja, wyszukiwanie multimodalne lub generowanie danych syntetycznych opartych na sygnale multimodalnym. Badamy szereg różnych metod głębokich sieci neuronowych do modelowania wspólnej przestrzeni multimodalnej i pokazujemy ich wyższość nad metodami pojedynczej modalności. Proponowane metody są niezależne od typu danych i mogą być stosowane w szerokim zakresie. Chociaż w większości skupiamy się na modalnościach obrazowych i tekstowych, uwzględniamy również eksperymenty z innymi modalnościami, takimi jak mimika i punkty charakterystycze twarzy.

Niniejsza rozprawa udowadnia następujące hipotezy badawcze:

- Metody głębokiego uczenia ulepszają przetwarzanie danych multimodalnych do zadań klasyfikacji i wyszukiwania.
- Modelowanie zależności kontekstowych zwiększa podobieństwo stylistyczne wyników wyszukiwania wielomodalnego.
- Zastosowanie metod generatywnych do danych multimodalnych prowadzi do uzyskania wyjaśnialnych reprezentacji wektorowych.

W pierwszej części rozprawy skupiamy się na metodach nadzorowanego głębokiego uczenia do uczenia reprezentacji multimodalnych. W rozdziałach czwartym i piątym wprowadzamy pomocnicze zadania optymalizacyjne i funkcje straty (takie jak zgodność stylów elementów czy strata trójkowa). W rozdziale szóstym wprowadzamy nowe generatywne metody uczenia wielomodalnych reprezentacji, które również posiadają własności wyjaśnialności. Rozszerzamy zapytanie multimodalne o syntetycznie wygenerowany obraz, który ilustruje informacje semantyczne zarówno z obrazu, jak i tekstu. Wreszcie, w rozdziale siódmym, przedstawiamy zastosowanie tych metod dla innego typu sygnału multimodalnego, aby pokazać szeroką gamę zastosowań proponowanych metod.

Metody głębokiego uczenia pozwoliły na wyuczenie reprezentacji wektorowych, opierając się na danych, bez potrzeby ręcznego tworzenia cech. Po tym, jak sieci neuronowe zostały z powodzeniem wykorzystane do uczenia osadzania obrazów i tekstu, zastosowano je również do danych multimodalnych (patrz: rozdział 2). W artykule Tautkute et al., 2017 przedstawiono metodę głębokiego uczenia dla multimodalnego uczenia osadzania, która jest ewaluowana na zadaniu wyszukiwania danych. Model nazwany Style Search Engine łączy unimodalną ekstrakcję cech z metodami mieszania cech. Metody mieszania łączą wyniki, które niezależnie maksymalizują podobieństwo wizualne, podobieństwo tekstowe i podobieństwo kontekstowe, gdzie podobieństwo kontekstowe jest empirycznym prawdopodobieństwem, że elementy pojawiają się razem w tym samym kontekście stylistycznym.

Proponowaną architekturę budujemy poprzez połączenie dwóch komponentów o pojedynczej modalności. Przeprowadzamy dwa wyszukiwania niezależnie dla zapytań obrazowych i tekstowych, w wyniku czego otrzymujemy dwa początkowe zestawy wyników. Następnie, wybieramy najlepsze dopasowania z początkowej puli wyników zgodnie z metodami mieszania. Po uzyskaniu wszystkich wizualnych i tekstowych dopasowań, algorytm mieszający szereguje je w zależności od podobieństwa w odpowiednich przestrzeniach cech i zwraca wynikową listę stylistycznie i estetycznie podobnych obiektów.

W artykule wprowadzono nową metrykę oceny podobieństwa stylów, która jest definiowana jako empiryczne prawdopodobieństwo pojawienia się elementów p_1, p_2 w tym samym, zdefiniowanym przez użytkownika zbiorze. Jest ona obliczana jako liczba zgodnych zbiorów p_1 i p_2, w których występują zarówno elementy p_1, jak i p_2, znormalizowana przez maksymalną liczbę zgodnych zbiorów, w których występuje którykolwiek z tych elementów. Metryka ta może być interpretowana jako empiryczne prawdopodobieństwo pojawienia się dwóch obiektów p_1 i p_2 w tym samym zbiorze zgodnym i jest wyrażana przez wynik podobieństwa leżący w przedziale $[0, 1]$.

$$s_c(p_1, p_2) = \frac{|\{C_i \in \mathcal{C} : p_1 \in C_i \wedge p_2 \in C_i\}|}{\max_{p \in \{p_1, p_2\}} \mid \{C_j \in \mathcal{C} : p \in C_j\} \mid}. \tag{1}$$

Łączenie multimodalnych wyników nie jest zadaniem trywialnym, dlatego też wprowadzamy kilka rodzajów metod mieszania i eksperymentalnie wybieramy najlepszą. Wprowadzona metoda zwiększa jakość wyników wyszukiwania i uzyskuje poprawę miary podobieństwa o 11%. Uśrednione wartości podobieństwa osiągają wartość $0,2484$ dla mieszanych cech głębokich, przewyższając proste metody mieszania, które dają wynik $0,2387$ średniego podobieństwa.

W artykule przedstawiono również nową bazę danych dla zadań wyszukiwania multimodalnego. Zbiór danych składa się z multimodalnych etykietowanych

danych z dziedziny wystroju wnętrz i może być wykorzystany w aplikacjach e-commerce. Produkty są opisane zarówno wizualnie jak i tekstowo i są skategoryzowane na kategorie produktów oraz zestawy kontekstowe. Zestaw kontekstowy jest stylistycznie i estetycznie spójnym zestawem mebli, które są umieszczone w tym samym zainscenizowanym pomieszczeniu przez projektantów.

Optymalizacja pomocniczych funkcji straty, takich jak funkcja starty do zadania klasyfikacji lub funkcja straty tripletowa (ang. *triplet loss*) poprawia proces uczenia się wielomodalnych reprezentacji. Dzięki dodatkowemu zadaniu optymalizacyjnemu, głęboka sieć neuronowa uczy się reprezentacji, zachowujących cechy i właściwości danych wejściowych.

Tripletowa funkcja straty (ang. *triplet loss*) i kontrastowa funkcja straty (ang. *contrastive loss*) są szczególnie użyteczne w odniesieniu do zadania uczenia metryk. Pomagają one nauczyć się przestrzeni osadzania, która ma pożądaną metrykę odległości. Proces uczenia następuje poprzez dostarczanie sieci przykładów bliskich i odległych punktów danych.

W artykule Tautkute et al., 2019 przedstawiono metodę, która wykorzystuje architekturę sieci neuronowych do modelowania wspólnej multimodalnej przestrzeni, z wykorzystaniem dodatkowych funkcji straty. Metoda ta jest rozszerzeniem poprzedniej pracy (Tautkute et al., 2017). W tym przypadku, podobieństwo stylów (aproksymowane przez empiryczne podobieństwo kontekstów) służy jako metryka odległości, której sieć się stara nauczyć. Dodatkowo, standardowa funkcja strata klasyfikacyjna służy jako pomocnicza funkcja straty , która pomaga zachować informacje semantyczne obecne w sygnale multimodalnym.

Po przeprowadzeniu szeregu eksperymentów w ostatecznej wersji modelu zastosowano sieć syjamską, w której każda gałąź posiada podwójne wejście składające się z cech obrazu i tekstu. Pozytywne pary są generowane jako pary obraz-tekst z tego samego zestawu kontekstowego, podczas gdy niepowiązane pary są uzyskiwane poprzez losowe próbkowanie elementów (obrazu i opisu tekstowego) z innego zestawu kontekstowego.

Przeprowadzona ewaluacja numeryczna pokazuje ulepszone możliwości wyszukiwania i przewyższa inne popularne metody bazowe do przetwarzania danych multimodalnych, takie jak MUTAN (Ben-younes et al., 2017) lub VSE (Kiros et al., 2014), przewyższając najsilniejszą metodę bazową (VSE-VGG19) o 21% dla zbioru danych dotyczących wystroju wnętrz. Proponowana metoda daje również najlepsze średnie wyniki podobieństwa dla zbioru danych dotyczących mody. Pod względem wartości dla obliczonej metryki jest ona o 32% wyższa, w porównaniu do najsilniejszego modelu bazowego.

Zaprezentowana metoda ma zastosowanie w domenach wystroju wnętrz i mody, jednak jej zastosowania nie ograniczają się do testowanych baz danych i z łatwością może być zastosowana w innych domenach zawierających dane multimodalne, takich jak pojazdy, artykuły produkcyjne czy dane dotyczące zdrowia.

Podczas gdy metody uczenia multimodalnego osadzania wprowadzone w poprzednich rozdziałach są w stanie zachować informacje semantyczne obecne w sygnale multimodalnym, nie oferują one wglądu w wyjaśnialność i interpretację wektorów ukrytych. W kolejnej części rozprawy doktorskiej zaproponowano możliwe rozwiązanie, które oferuje metodę uczenia wyjaśnialnych reprezentacji wektorowych.

W tym celu proponowana jest metoda oparta na GAN, którą nazywamy SynthTriplet GAN. Generuje ona syntetyczny obraz, który odpowiada informacji semantycznej zawartej w zapytaniu i może uprościć multimodalny proces wyszukiwania do bezpośredniego wyszukiwania wizualnego. Zostało to osiągnięte poprzez

wprowadzenie nowej metody wyboru tripletów w funkcji straty tipletowej, z syntetycznie wygenerowanym elementem odniesienia (ang. *anchor*) w celu optymalizacji nauczonych reprezentacji wektorowych dla zadania wyszukiwania.

W szczególności, podczas treningu dostarczane są dwa obrazy - obraz wejściowy i pożądany obraz docelowy. Kotwica jest generowana syntetycznie na podstawie obrazu wejściowego i zapytania tekstowego $\hat{t}$. Stąd, dla obrazu wejściowego x, obrazu docelowego $\hat{x}$ względnego zapytania tekstowego $\hat{t}$, zmodyfikowana strata trójkowa jest następująca:

$$L_{tr}(x, \hat{x}, \hat{t}) = max(0, d(\hat{y}, \hat{x}) - d(\hat{y}, x) + m) \tag{2}$$

gdzie $d(x, y)$ to odległość między zakodowanymi wektorami cech, a m to margines. W tym scenariuszu obrazem kotwiczącym jest wygenerowany obraz $\hat{y}$, obrazem pozytywnym jest obraz docelowy $\hat{x}$, a obrazem negatywnym jest obraz źródłowy x.

Ponadto, niniejsza rozprawa udowadnia, że wyuczone reprezentacje metodą SynthTriplet GAN zapewniają dodatkową interpretowalność, w przeciwieństwie do niegeneratywnych metod multimodalnych. Ilustrujemy, że poprzez liniową interpolację pomiędzy dwoma sygnałami tekstowymi wejściowymi dla tego samego obrazu źródłowego, możemy osiągnąć stopniową zmianę w syntetycznie generowanym obrazie, która zachowuje informacje semantyczne. Na przykład, interpolując liniowo pomiędzy wektorami tekstowymi reprezentującymi kolory niebieski i żółty oraz utrzymując obraz wejściowy na stałym poziomie, otrzymujemy obrazy pośrednie w kolorach zielonych (zgodnie z oczekiwaniami teorii kolorów).

W artykule Tautkute and Trzcinski, 2021 przedstawiono ilościową ocenę wykorzystania obrazów syntetycznych w multimodalnym wyszukiwaniu i wykazano znaczącą poprawę w stosunku do innych metod generatywnych. Analiza studium ablacji pokazuje wpływ użycia strat trójkowych na wartości przywoływania i ich znaczenie w multimodalnym wyszukiwaniu. Jest to jedna z pierwszych prac, która dostarcza tego rodzaju analizę i proponuje nowe zastosowanie metod generatywnych.

W ostatniej części rozprawy poruszono temat innych typów sygnałów multimodalnych. Chociaż zakres wszystkich możliwych modalności wykracza poza zakres rozprawy, pokazano, że wspomniane metody są możliwe do zastosowania i rozszerzenia również na inne dziedziny i typy sygnałów. Przykładem takiego zastosowania jest dodanie modalności położenia punktów orientacyjnych twarzy do zadania rozpoznawania ekspresji twarzy i emocji. Lokalizowanie punktów orientacyjnych twarzy i rozpoznawanie emocji jest ważną dziedziną interakcji człowiek-komputer i jest niezbędne dla rozwoju sztucznej inteligencji. Pokazujemy, że samo rozszerzenie problemu o nową modalność i zastosowanie uczenia wielozadaniowego poprawia dokładność predykcji systemów rozpoznawania wyrazu twarzy.

W artykule Tautkute and Trzcinski, 2019 zaprezentowano, że poprzez rozszerzenie jednomodalnego zadania klasyfikacji emocji z ludzkich twarzy o dodatkową modalność informacyjną i zadanie lokalizacji punktów orientacyjnych twarzy, poprawiono wyniki w stosunku do pojedynczej modalności. Ponadto pokazano, że dodatkowa modalność powoduje rozpoznanie ważnych regionów ludzkiej twarzy przez sieć, skorelowanych z udokumentowanymi jednostkami akcji (AU) dla ekspresji emocji (punktami orientacyjnymi twarzy odpowiedzialnymi za wyrażanie emocji).

Zaprezentowana metoda działa w środowisku wielozadaniowym poprzez rozszerzenie funkcji straty architektury DAN (Kowalski et al., 2017) o funkcję straty entropii krzyżowej (wyjście softmax dla przewidywanej kategorii emocji). Oba pojęcia

są optymalizowane podczas procesu treningowego, a parametry wagowe strat są ustalane empirycznie.

W części eksperymentalnej przedstawiono ocenę numeryczną na kilku zbiorach danych dotyczących rozpoznawania wyrazu twarzy: AffectNet (Mollahosseini et al., 2017), CK+ (Lucey et al., 2010), JAFFE (Lyons et al., 1998), ISED (Happy et al., 2017) i pokazano poprawioną dokładność klasyfikacji $F1 - score$ w porównaniu do metod bazowych.

Ponadto, dokonano wizualnego objaśnienia decyzji klasyfikacyjnych, przy użyciu gradientowej techniki lokalizacyjnej Grad-CAM (Selvaraju et al., 2017) i pokazano, że większość decyzji związanych z klasyfikacją emocji jest oceniana na podstawie regionów wokół ust, oczu, nosa i brwi. Model poprawnie identyfikuje te regiony, mimo że sieć nie dostaje informacji lokalizacyjnych.

Podsumowując, w tej rozprawie przedstawiono nowe nadzorowane i generatywne metody przetwarzania danych multimodalnych, które otwierają nowe możliwości dla zadań takich jak klasyfikacja, wyszukiwanie danych multimodalnych czy generowanie obrazów syntetycznych. Pokazano, że przetwarzanie danych multimodalnych ma wiele zastosowań i jest niezwykle istotne, aby opracować dobre reprezentacje danych multimodalnych w celu zbudowania inteligentnych systemów przyszłości. Modelowanie reprezentacji multimodalnych, które zachowują informacje semantyczne, jest podstawowym zadaniem dla efektywnych zastosowań uczenia maszynowego. Pomyślny rozwój takich metod poprawi zdolność maszyn do efektywnego poruszania się po ogromnych zbiorach danych i otworzy nowe ścieżki w rozumieniu danych przez systemy komputerowe.