Prof. dr hab. inż. Bożena Kostek, czł. koresp. PAN
Politechnika Gdańska,
Wydział Elektroniki, Telekomunikacji i Informatyki
Lab. Akustyki Fonicznej

28.05.2022 r.

Opinia nt. rozprawy doktorskiej mgr inż. Ivony Tautkute

pt.: „**Artificial Neural Networks for Multimodal Data Embeddings and Classification**", wykonanej pod kierunkiem dr hab. Alicji Wieczorkowskiej oraz dra hab. inż. Tomasza Trzcińskiego, prof. PW.

1. Jakie zagadnienie naukowe/badawcze jest rozpatrywane w pracy (cel i teza rozprawy) i czy zostało ono dostatecznie sformułowane przez autorkę

Przedmiotem recenzji jest rozprawa doktorska mgr inż. Ivony TAUTKUTE pt.: "**Artificial Neural Networks for Multimodal Data Embeddings and Classification**". Głównym zagadnieniem badawczym jest zaprojektowanie algorytmów uczenia maszynowego, które pozwoliłyby na inteligentne przetwarzanie/wyszukiwanie danych łączących różne modalności. W szczególności dotyczy to problemu uczenia multimodalnych reprezentacji wektorowych z jednoczesnym modelowaniem użytecznych reprezentacji, zachowujących kontekst semantyczny z wielu kanałów informacyjnych.

Recenzowana rozprawa doktorska ma charakter teoretyczno-eksperymentalny, składa się z ośmiu rozdziałów: wprowadzenia, przeglądu prac związanych z omawianymi w pracy zagadnieniami, zbiorczego podsumowania opublikowanych prac, stanowiące rdzeń niniejszej rozprawy i które zostały zawarte w rozdziałach 4-7, rozdziału podsumowującego oraz Bibliografii. W pracy znajdują się również streszczenia w j. angielskim i polskim, lista rysunków oraz wykaz skrótów, brakuje natomiast wykazu wielkości matematycznych. Rozprawa zawiera się w 115 stronach tekstu wraz z dodatkami.

We Wprowadzeniu doktorantka podaje w pierwszej kolejności genezę oraz motywację prowadzonych prac badawczych odnoszących się do szeroko sformułowanej tematyki inteligentnego przetwarzania/wyszukiwania danych multimodalnych. Tematyka pracy obejmuje opracowanie wyjaśnialnych zanurzeń/osadzeń (ang. *embeddings*) danych multimodalnych, znajdujących zastosowanie w klasyfikacji, wyszukiwaniu wielomodalnym, generowaniu obrazów syntetycznych na podstawie danych wielomodalnych.

Autorka rozprawy sformułowała trzy hipotezy badawcze:

(w j. ang.):

1. **Deep learning methods improve multimodal data processing for classification and retrieval tasks.**
2. **Modeling contextual dependencies increases stylistic similarity of the multimodal retrieval results.**
3. **Applying generative methods to multimodal data results in explainable embeddings.**

(w j. polskim)

1. **Metody głębokiego uczenia usprawniają przetwarzanie danych multimodalnych do zadań klasyfikacji i wyszukiwania.**
2. **Modelowanie zależności kontekstowych zwiększa podobieństwo stylistyczne multimodalnych wyników wyszukiwania.**
3. **Zastosowanie metod generatywnych do danych multimodalnych prowadzi do powstania wyjaśnialnych (wytłumaczalnych) zanurzeń/osadzeń (ang. *embeddings*).**

W końcowej części Wprowadzenia zostały wylistowane osiągnięcia Autorki rozprawy, będące wynikiem prowadzonych prac badawczych w ramach rozprawy doktorskiej.

2. **Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym, literatury światowej, stanu wiedzy i zastosowań w przemyśle**

Analiza literatury światowej została zawarta w rozdziale 2. Lista źródeł zawiera ok. 120 pozycji i została uszeregowana w kolejności cytowania. Zawiera źródła podane w przeglądzie literatury (rozdz. 2), jak i w publikacjach stanowiących rozdziały 4-7. Pewien niedosyt budzą nieliczne cytowania do źródeł z roku 2020 oraz 2021. Sądzę, że warto by było sięgnąć do aktualnie prowadzonych prac/przeglądu dokonań w tematyce rozprawy doktorskiej (np. Bayoudh, K., Knani, R., Hamdaoui, F. et al., A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. Vis. Comput., 2021; https://doi.org/10.1007/s00371-021-02166-7 czy Wei Chen, Weiping Wang, Li Liu, Michael S. Lew, New Ideas and Trends in Deep Multimodal Content Understanding: A Review, Neurocomputing, vol. 426,2021, pp. 195-215, https://doi.org/10.1016/j.neucom.2020.10.042).

W genezie rozprawy doktorskiej, przeglądzie prac, jak również w zawartych publikacjach pojawiają się odniesienia do zastosowań praktycznych (i w przemyśle), co jest istotne z punktu widzenia praktycznego wykorzystania uzyskanych wyników

badań. W ogólności prowadzone eksperymenty są ukierunkowane na zastosowania w praktyce.

3. **Czy autorka rozwiązała postawione zagadnienia; czy użyła właściwej do tego metody i czy przyjęte założenia są uzasadnione**

Przyjęta metodologia – warunkowana hipotezą badawczą jest w pełni poprawna i odnosi się do stanu wiedzy. W ramach rozprawy wykorzystywany jest szeroki przekrój modeli głębokich, jak również miar strat/jakości proponowanych rozwiązań, adaptowanych na podstawie literatury, ale odnoszących się do prezentowanych problemów. Warto również zwrócić uwagę, że Autorka rozprawy zaproponowała szereg podejść metodologicznych, które pozwalają na stwierdzenie, że rozprawa doktorska ma charakter nie tylko eksperymentalny, ale również teoretyczny.

4. **Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy i poziomu techniki reprezentowanych przez literaturę światową**

Prezentowana tematyka jest aktualna i ważna w świetle stanu wiedzy. Widoczne są liczne odniesienia w literaturze – z jednej strony – w kontekście konsolidowania różnych modalności, zwłaszcza w odniesieniu do praktycznych zastosowań, zaś z drugiej – poszukiwania algorytmów inteligentnego przetwarzania tego typu danych. Przy czym warto zauważyć, że nie jest to tworzenie rozwiązań dla „sztucznych" problemów, gdyż zasoby dostępne w Internecie występują najczęściej w postaci złożonej z dwóch i więcej modalności. Autorka rozprawy podaje następujące osiągnięcia uzyskane w trakcie badań, które dotyczą propozycji metodologii i ich praktycznej implementacji:

1. Konstrukcja głębokiej sieci neuronowej (*Style Search Engine*) dla multimodalnej wyszukiwarki, która łączy wyniki dla różnych modalności. Proponowana metoda integruje wykrywanie obiektów, wyszukiwanie wizualne i tekstowe z podobieństwem stylów pobranych obiektów, zwiększając jakość uzyskanych wyników.

Blok wyszukiwania wizualnego wykorzystuje wykrywanie obiektów i dane wyjściowe wstępnie wytrenowanej sieci CNN. Blok tekstowy pozwala na sprecyzowanie kryteriów wyszukiwania za pomocą tekstu i zwiększa kontekstowe znaczenie wyszukanych wyników. Wreszcie, poprzez łączenie wizualnych i tekstowych wyników wyszukiwania za pomocą wyników podobieństwa w przestrzeniach cech, metoda ta znacząco poprawia stylistyczne i jednocześnie estetyczne podobieństwo wyszukanych elementów.

2. Konstrukcja głębokiej splotowej syjamskiej sieci neuronowej, która łączy wizualne i tekstowe cechy obiektu. Wspólna architektura sieci neuronowych (DeepStyle) modeluje kontekstowo zależności między cechami różnych modalności i pozwala na ich wyszukiwanie poprzez obie modalności (obraz i tekst).

Warto zauważyć, że w wielu zastosowaniach wyszukiwania multimodalnego, trudno jest jednoznacznie określić miarę potrzeb użytkownika i motywację do zakupu danego produktu. W literaturze – na co zwraca uwagę Autorka rozprawy – zdefiniowano wiele pomocniczych metryk, które przybliżają inne aspekty podobieństwa wyników wyszukiwania, jak na przykład ich zgodności stylistycznej. Nawet, jeśli tego typu metryka jest trudna do zdefiniowania, to istnieją w literaturze przykłady takich definicji. Autorka rozprawy, wykorzystuje formułę Ferniego, w której styl jest definiowany jako sposób, który pozwala grupować obiekty w powiązane kategorie, a następnie modyfikuje tę definicję, stosując podejście probabilistyczne do wyznaczania miary podobieństwa stylistycznego. Tego typu podejście nie wymaga stosowania ręcznie tworzonych cech i predefiniowanych stylów. W uproszczeniu można powiedzieć, że proponowana metoda jest wzmacniana miarą podobieństwa, co jest istotne w szczególności w przypadku słabej relacji kontekstowej występujących obiektów. Dodatkowo p. Ivona Tautkute rozszerza metodologię oceny podobieństwa za pomocą syjamskiej głębokiej sieci neuronowej (DeepStyle).

3. Konstrukcja modelu (SynthTriplet GAN) wykorzystującego generatywne sieci współzawodniczące (ang. *Deep Generative Adversarial Network*) z wejściem multimodalnym. Model zapewnia wytłumaczalne multimodalne osadzenia z wykorzystaniem zapytań w postaci syntetycznie generowanych obrazów.

Proponowany model głęboki (architektura SynthTriplet GAN), składający się na wejściu z enkodera obrazu i tekstu, rozszerza zapytanie multimodalne o syntetycznie wygenerowany obraz za pomocą generatora, który przechwytuje informacje semantyczne pochodzące zarówno z obrazu, jak i tekstu. Autorka wprowadza w ten sposób nową metodę eksploracji, która wykorzystuje syntetyczny obraz jako element bezpośredniej optymalizacji odległości osadzania obrazów generowanych i docelowych – rzeczywistych. Miara straty określona jest jako triplet strat, aby zapewnić jak najmniejszą odległość między syntetycznymi obrazami wygenerowanymi na podstawie zapytania multimodalnego a obrazem rzeczywistym. W metodzie tej na wejście podawane są dwa obrazy, zaś kotwicę stanowi syntetycznie wygenerowany obraz, tworzony ze źródłowego zapytania bimodalnego (obraz i tekst). Istotne jest też porównanie uzyskanych wyników z metodami stosowanymi przez innych badaczy, które wskazują na wyższość proponowanej metodologii.

4. Konstrukcja sieci splotowych (EmotionalDAN) w podejściu do multimodalnej klasyfikacji emocji na podstawie łącznej oceny punktów orientacyjnych twarzy zmapowanych na płaszczyźnie 2D oraz obrazu twarzy zawierającego ekspresję emocji twarzy.

Wykorzystanie zaproponowanych modeli głębokich w treningu sygnałów multimodalnych innego typu, jak punkty orientacyjne rozmieszczone na twarzy oraz mimika twarzy, powoduje znaczny wzrost dokładności klasyfikacji (5% w stosunku do stanu wiedzy). Można zauważyć, że przeprowadzone eksperymenty ilustrują poprawę zdolności generalizacji zaproponowanych modeli i metod. Stanowi to rozszerzenie proponowanej metody uczenia osadzeń multimodalnych poprzez łączną optymalizację straty podobieństwa stylów wraz ze stratą klasyfikacji. Warto zauważyć, że i w tym przypadku Autorka rozprawy wykorzystuje opracowaną funkcję strat

Podsumowując osiągnięcia Autorki rozprawy, wydaje mi się, że niezwykle istotnym aspektem rozprawy jest opracowanie (zaadoptowanie z literatury) miar oceny jakości działania/zgodności stylistycznej proponowanych modeli głębokich.

Pewien niedosyt budzi jedynie brak szczegółów dotyczących budowy/konstrukcji baz danych. Nie jest wiadome, ile rekordów zawiera/ją ta/te bazy, w jaki sposób uzyskiwano dane, w jaki sposób przebiegał proces adnotacji oraz weryfikacji tego procesu, itp. W rozdziale 3. można znaleźć odniesienia do weryfikacji zapytań przez użytkownika, ale nie wyjaśnia to w pełni tego aspektu.

W tym kontekście nasuwa się też pytanie, w jakim stopniu skonstruowane modele są skalowalne oraz czy można je wykorzystać w ujęciu do innego zestawu modalności? Druga część tego pytania stanowi element do dyskusji.

Podsumowując, osiągnięciem Autorki rozprawy doktorskiej są przygotowane struktury algorytmów, jak również opracowanie metodologii treningu modeli głębokich, wyszukiwania/przyporządkowania poszczególnych modalności, co skutkuje udowodnieniem hipotez badawczych.

5. **<u>Czy autorka wykazała umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?</u>**

Rozprawa doktorska jest przedstawiona w sposób logiczny i przekonywujący. Ze względu na konstrukcję rozprawa niewątpliwie ma zwięzły charakter, zawiera wprowadzenie do tematu, hipotezy badawcze, odniesienie do stanu wiedzy i następnie opublikowane prace, które stanowią rdzeń rozprawy. Praca jest poprawna od strony redakcyjnej, edycja rozprawy jest staranna. Przedstawione w rozdziale 3. schematy

blokowe stanowią czytelną ilustrację prowadzonych badań i stanowią niewątpliwie wartość dodaną. Ponadto, krótkie wprowadzenia poprzedzające zawarte w rozprawie publikacje (rozdz. 4-7) zapewniają ciągłość przedstawianych wątków badawczych. Jednak – jak wspomniano wcześniej – brakuje bardziej szczegółowego przedstawienia sposobu tworzenia baz danych.

6. **Jaka jest przydatność rozprawy dla dyscypliny?**

Tematyka zawarta w rozprawie jest ważna i istotna w każdym przedstawianym w rozprawie kontekście. Aktualność tej tematyki jest potwierdzona poprzez aktualnie prowadzone badania i publikacje. Prowadzone badania wpisują się dobrze w dyscyplinę, w której doktoryzuje się p. Ivona Tautkute.

Podsumowanie

W podsumowaniu stwierdzam, że przedłożona mi do recenzji rozprawa doktorska p. Ivony Tautkute **spełnia z nadmiarem wymagania** stawiane rozprawom doktorskim w aktualnie obowiązującej Ustawie o stopniach i tytule naukowym. W związku z tym wnoszę **o dopuszczenie rozprawy doktorskiej p. mgr inż. Ivony Tautkute do publicznej obrony.**

Ze względu na osiągnięte w rozprawie wyniki, które zostały opublikowane w wysoko punktowanych czasopismach (zgodnie z warunkami przyznawania wyróżnień w PJATK jedna z prac ukazała się na konferencji 28th International Conference on Neural Information Processing (ICONIP2021), która ma przypisane 140 punktów wg listy MEiN. Pani Ivona Tautkute jest pierwszą autorką tej pracy. Praca stanowi część rozprawy doktorskiej), **wnoszę wyróżnienie rozprawy.**

Bricue Kod