Częstochowa, dn. 10 lutego 2022 r.

prof. dr hab. inż. Rafał Scherer
Katedra Inteligentnych Systemów Informatycznych
Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska
al. Armii Krajowej 36
42-200 Częstochowa

Recenzja

rozprawy doktorskiej mgra inż. Mateusza Modrzejewskiego, pt.: Artificial Intelligence Solutions for Artistic Multimedia Musical Content Creation Support.

Promotor: prof. dr hab. inż. Przemysław Rokita

Niniejszą recenzję opracowano na wniosek Rady Dyscypliny Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej, która mocą uchwały z dnia 23 listopada 2021 roku powołała mnie na recenzenta.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Po powolnym rozwoju i tak zwanych zimach, doświadczamy obecnie niespotykanego rozwoju, wręcz eksplozji, metod uczenia maszynowego. Są one związane głównie z głębokimi sieciami neuronowymi. Metody oparte o nie wyszły z laboratoriów i są stosowane do rozwiązywania szeregu problemów, nierozwiązywalnych wcześniej lub takich, które napotkały barierę w przypadku użycia metod tradycyjnych. Przykładami są tu autonomiczne pojazdy, przetwarzanie oraz generowanie tekstu i multimediów, automatyczna translacja miedzy językami, a nawet dowolną zawartością, automatyczna gra w typowo ludzkie gry. W ostatnich latach, we wszystkich tych zadaniach zazwyczaj sztuczne sieci neuronowe pokonują człowieka. Niektóre z tych obszarów, szczególnie związane ze sztuką, są wysoce subiektywne w ocenie. Taki również jest obszar zagadnień, którego dotyczy oceniana rozprawa, a mianowicie przetwarzanie i generowanie muzyki. Tak jak ciężko ocenić w sposób analityczny zawartość treści wizualnych, tak samo trudno ocenić muzykę. Celem rozprawy było stworzenie metod transferu brzemienia pomiędzy utworami, generowania muzyki oraz klasyfikacji muzyki ze względu na jej gatunek.

2. Zawartość rozprawy

Recenzowana praca mgra inż. Mateusza Modrzejewskiego składa się z sześciu rozdziałów, dodatków, bibliografii oraz spisu rysunków i tabel. Dokument liczy 135 stron.

Pierwszy rozdział jest krótkim wprowadzeniem do głębokiego uczenia. Dalej wymienione jest pięć celów szczegółowych rozprawy oraz krótko opisano strukturę dokumentu.

Rozdział 2 jest krótkim wprowadzeniem do teorii muzyki. Omówiono rytm, melodię, harmonię i strukturę. Dalej omawiane są sposoby reprezentacji muzyki, takie jak pięciolinia i partytura, format MIDI i ABC. Dalej przedstawione są cechy wyliczane, takie jak obwiednia amplitudy, energia średniokwadratowa, współczynnik przechodzenia przez zero, stosunek energii pasm, centroidę spektrum, szerokość spektrum, strumień spektrum, początek nuty, spektrogramy, chromagramy, centroidy tonalne czy współczynniki MFCC. Reprezentacja MIDI oraz zaawansowane wyliczeniowe cechy sygnału muzycznego będą użyte w następnych rozdziałach jako reprezentacja danych dla sieci neuronowych.

Rozdział 3 dotyczy idei transferu stylu, czyli przeniesienia (nałożenia) danego stylu artystycznego na zawartość stworzoną w innym stylu. Opisano początkowe prace Gatys'a i innych z 2015 dotyczące transferu stylu w plikach graficznych. Prace te wykazały możliwość odseparowanej reprezentacji stylu i zawartości w neuronowych sieciach splotowych. Dalsza część rozdziału dotyczy transferu stylu w muzyce z naciskiem na transfer brzmienia z podziałem na użycie autoenkoderów, autoenkoderów wariacyjnych, oraz ich modyfikacji jak MoVE czy Universal Music Translation Network.

Dalej przedstawiono autorskie rozwiązanie, którego pierwotna wersja została opublikowana na konferencji Artificial Intelligence and Soft Computing w 2021 roku. Metoda była opracowana z użyciem dwóch dużych zbiorów danych MIDI. Z nich Autor wybrał utwory, których przeważająca częścią był pojedynczy instrument, a następnie syntezował za pomocą programowego syntezatora muzykę. Danymi wejściowymi do sieci neuronowych był spektrogram STFT oddzielnie w dziedzinie rzeczywistej i urojonej.

Pierwszym modelem neuronowym zaproponowanym przez Doktoranta jest Baseline autoencoder T0, składający się z warstw typu nieliniowy perceptron. Liczba neuronów w warstwie wyjściowej odpowiada liczbie współczynników transformaty STFT. Niejasne może być dla czytelnika dlaczego występują dwie takie warstwy. W następnym modelu – Recurrent autoencoder TLSTM1, mamy oddzielne warstwy wyjściowe dla części rzeczywistej i urojonej. Doktorant dodał do modelu dwie ukryte warstwy LSTM. Następne modele z warstwami LSTM również mają warstwy wyjściowe dla części rzeczywistej i urojonej, ale podwojone. Ponadto, dane wejściowe wchodzą bezpośrednio na warstwę LSTM. Ostatni model, TLSTM5, jest jeszcze bardziej uproszczony co spowodowało jeszcze mniejsze zapotrzebowanie na pamięć. Eksperymenty na syntezowanych utworach MIDI pokazały, że każdy następny model dawał lepszą dokładność translacji. Sekwencyjna natura sieci LSTM dała oczywiste polepszenie jakości translacji i słyszalnie lepszą jakość generowanego dźwięku.

W rozdziale 4 Doktorant podejmuje tematykę automatycznego generowania (komponowania) muzyki. Omówiono kilka rozwiązań istniejących w literaturze. Autor przedstawił swoje autorskie rozwiązanie oparte o sieci GAN, którego pierwotna wersja została opublikowana na konferencji Artificial Intelligence and Soft Computing w 2019 roku. Do trenowania sieci GAN zostały użyte cztery zbiory muzyki MIDI. Dane zostały skonwertowane najpierw do postaci tekstowej i usunięto wszelkie niepotrzebne znaczniki MIDI oraz zawężono zakres do 64

wysokości dźwięków, przesuwając skrajne dźwięki o oktawę. Doktorant usunął również wszelkie ewentualne błędy w danych, takie jak, na przykład, niezakończone, lub zakończone, a nie otwarte polecenia MIDI. Tak przygotowane dane zostały zamienione na obrazy RGB o rozdzielczości 64×64 w formacie piano roll (taśma pianoli). Przez dodatkowe kodowanie nut kolorem można było zmieścić 20 sekund muzyki z takim obrazie. Tak przygotowane obrazy posłużyły do trenowania sieci GAN składającej się z pełni splotowego generatora i dyskryminatora. Autor wykonał wiele eksperymentów i dokonał oceny wizualnej wygenerowanych obrazów oraz słuchowej, analizując harmonię oraz rytm. Jak już wspomniałem, obiektywna względna lub bezwzględna ocena treści generowanych przez sieci GAN jest bardzo trudna. Ciekawostką jest użycie wygenerowanych próbek MIDI do stworzenia mini albumu przez profesjonalnego muzyka z użyciem oprogramowania Ableton Live.

Rozdział 5 dotyczy metod klasyfikacji muzyki ze względu na jej gatunek. Zaprezentowano krótki przegląd literaturowy dotyczący tematyki i istniejących zbiorów danych. Autorska metoda doktoranta jest rozwinięciem pierwotnej wersji opublikowanej na konferencji Artificial Intelligence and Soft Computing w 2019 roku. Autor użył istniejącego zbioru oznaczonych plików muzycznych, z których wyekstrahowano spektrogramy i chromagramy. Autor dokonał ciekawego przeglądu popularnych gatunków muzycznych pod kątem cech widocznych na spektrogramach i chromagramach. Pierwszymi zaproponowanymi modelami są typowe sieci splotowe składające się z warstw splotowych i MLP (gęstych, w pełni połączonych) do ostatecznej klasyfikacji. Model 1 posiada dwie warstwy splotowe, a model 2 – cztery. Trzecim modelem jest równoległa sieć splotowo-rekurencyjna. W tym zadaniu można było już oczywiście podać odpowiednie wskaźniki jakości klasyfikacji. Model 3 uzyskał najlepsze wyniki, porównywalne do tych z literatury. Doktorant dokonał również ciekawej analizy wyników oceniając wyniki pod kątem specyfiki klasyfikowanych gatunków muzycznych. W rozdziale brakuje dokładnych wymiarów danych wejściowych do sieci oraz porównania złożoności modeli z tymi występującymi w literaturze.

Rozdział 6 jest podsumowaniem rozprawy, w którym zebrano w jednym miejscu wnioski kończące opisy poszczególnych autorskich metod i eksperymentów. Doktorant podał również wiele pomysłów, które wytyczają wiele kierunków badawczych, np. użycie nowopowstałych architektur, np. CycleGAN, standardu MIDI 2.0 czy powstających danych tzw. odcisków palca utworów.

Dalej następują dodatki. Dodatek A będący zestawieniem publikacji Doktoranta. Dodatek B, w którym wymienia swoje imponujące osiągnięcia muzyczne. Dodatek C, w którym zebrał pewne informacje dotyczące legalności i praw autorskich dla muzyki generowanej automatycznie i używania utworów do trenowania sieci neuronowych.

Pracę kończy bardzo szeroka bibliografia składająca się z aktualnych pozycji, spis rysunków oraz tabel.

Ogólnie, zasadnicze i oryginalne rezultaty pracy można podsumować następująco:

- Opracowanie wprowadzenie do tematyki i dokonał przeglądu literatury dotyczącej automatycznej translacji, generowanie oraz klasyfikacji muzyki.

- Zaprojektowanie metody translacji brzmienia muzyki. Eksperymenty były przeprowadzone dla translacji brzmienia fortepianu na gitarę, ale metoda może być użyta dla dowolnych kombinacji, zależnie od przygotowanych danych.
- Stworzenie metody generacji muzyki za pomocą sieci GAN i danych MIDI przekształconych do formatu piano roll.
- Opracowanie metodę klasyfikacji gatunków muzycznych o różne modele sieci głębokich.
- Przeprowadzenie eksperymentów wraz z ciekawą analizą.

Mgr Modrzejewski opublikował osiem prac naukowych w materiałach konferencji. Zaprezentowany materiał pokazuje, że Doktorant zrealizował cel pracy.

3. Uwagi krytyczne i wskazówki dotyczące rozprawy

Praca napisana jest schludnie i przejrzyście. Praca obfituje w czytelne rysunki oraz schematy. Ponadto na uwagę zasługuje użycie języka angielskiego na bardzo dobrym poziomie. Poniżej zamieszczam kilka pytań, które zrodziły się w czasie czytania pracy:

Często brak dokładnych meta parametrów używanych modeli neuronowych. Np. dla modeli z rozdziału 5 nie podano rozmiarów danych wejściowych oraz rozmiarów filtrów splotowych.

Dlaczego Baseline autodencoder T0 z podrozdziału 3.3.2 ma rozmiar przestrzeni ukrytej 256. Czy było to wynikiem doświadczeń?

Dlaczego rozmiar wejściowy generatora w sieci GAN na rysunku 4.3 wynosi 100?

W literaturze funkcjonują rozwiązania problemów podejmowanych w pracy doktorskiej. Czy można by porównać złożoność modeli zaproponowanych w pracy z tymi z literatury? Czy można by porównać jakość klasyfikacji gatunków muzycznych z rozdziału 6 z wynikami osiąganymi w literaturze?

Nie ma informacji o sposobie implementacji zaproponowanych modeli.

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że Pan mgr inż. Mateusz Modrzejewski w rozprawie doktorskiej „ Artificial Intelligence Solutions for Artistic Multimedia Musical Content Creation Support":

- Zrealizował cel rozprawy,
- Opracował wprowadzenie do tematyki i dokonał przeglądu literatury dotyczącej automatycznej translacji, generowanie oraz klasyfikacji muzyki.
- Zaprojektował metodę translacji brzmienia muzyki. Eksperymenty były przeprowadzone dla translacji brzmienia fortepianu na gitarę, ale metoda może być użyta dla dowolnych kombinacji, zależnie od przygotowanych danych.

- Stworzył metodę generacji muzyki za pomocą sieci GAN i danych MIDI przekształconych do formatu piano roll.
- Opracował metodę klasyfikacji gatunków muzycznych o różne modele sieci głębokich.
- Wykazał się umiejętnością samodzielnej pracy badawczej, znajomością literatury światowej i wiedzą w zakresie uczenia maszynowego.
- Zadbał o popularyzację wyników swoich badań w materiałach konferencji międzynarodowych.

Recenzowana praca spełnia wymagania ustawy o tytule i stopniach naukowych w dyscyplinie naukowej Informatyka Techniczna i Telekomunikacja. Rozprawa doktorska prezentuje ogólną wiedze teoretyczną Doktoranta w dyscyplinie. Przedmiotem rozprawy jest oryginalne rozwiązanie problemu naukowego. Należy podkreślić ponadprzeciętną wiedzę, doświadczenie i osiągnięcia muzyczne Doktoranta. Wnoszę o jej przyjęcie i dopuszczenie do dalszych etapów postepowania doktorskiego.