Določanje podžanrov elektronske glasbe z uporabo M2D embeddingov

Anonimni Študent
Faculty of Computer and Information Science
University of Ljubljana
Email: xx0000@student.uni-lj.si

***Povzetek*—V članku predstavljamo razvoj in implementacijo sistema za klasifikacijo elektronske glasbe v podžanre. Sistem temelji na prednaučenem M2D modelu korporacije NTT, ki služi kot osnova za pridobivanje kompleksnih zvočnih reprezentacij. Primerjava z implementacijo konvolucijske nevronske mreže (CNN) je pokazala, da prednaučene M2D reprezentacije omogočajo bistveno boljšo točnost klasifikacije (76.0% proti 61.5% na testni množici). Po začetni zasnovi s kompleksnejšim Jukebox modelom smo zaradi računskih zahtev prešli na učinkovitejši M2D model, ki omogoča praktično uporabo na standardni strojni opremi. Na pridobljenih reprezentacijah smo razvili dodatni linearni klasifikator za razlikovanje med petimi žanri elektronske glasbe: ambient, drum and bass, house, techno in trance. Razvita rešitev vključuje celovit cevovod za procesiranje zvoka, učenje modela in produkcijsko uporabo, sestavljen iz učnega okolja, inference, REST API strežnika in spletnega vmesnika za končne uporabnike. Vmesnik omogoča enostavno klasifikacijo z vizualizacijo zaupanja v napovedi preko interaktivnega stolpičnega diagrama. Posebno pozornost smo namenili optimizaciji sistema za učinkovito delovanje v realnem času in analizi časovne ter prostorske zahtevnosti implementacije.**

I. Uvod

Klasifikacija elektronske glasbe v podžanre predstavlja kompleksen problem zaradi subtilnih razlik med žanri in pogosto nejasnih mej med njimi. Tradicionalni pristopi, ki temeljijo na preprostih zvočnih značilkah ali spektrogramih, pogosto ne zajamejo vseh pomembnih aspektov, ki določajo žanr. Z razvojem globokega učenja so se pojavili novi pristopi za reprezentacijo glasbe, med katerimi izstopata Jukebox [6] in M2D [**?**] model, ki omogočata zajem kompleksnih zvočnih značilnosti.

V tem delu predstavljamo implementacijo sistema za klasifikacijo elektronske glasbe, ki temelji na prednaučenem M2D modelu. Ključni prispevki našega dela so:

- Razvoj celotnega cevovoda za procesiranje zvočnih posnetkov v skladu z M2D specifikacijami
- Implementacija in optimizacija sistema za učenje linearnega klasifikatorja na M2D embeddingih
- Razvoj praktičnega spletnega vmesnika za klasifikacijo z vizualizacijo rezultatov
- Analiza časovne in prostorske zahtevnosti sistema

II. Sorodna Dela

Področje klasifikacije elektronske glasbe je doživelo več razvojnih faz. Zgodnji pristopi so temeljili na ročno zasnovanih značilkah kot so MFCC koeficienti [7], spektralni centroid in drugi spektralni deskriptorji. Z razvojem globokega učenja so se pojavili pristopi, ki temeljijo na konvolucijskih nevronskih mrežah [8] in transformerjih [9].

Jukebox model [6] predstavlja mejnik v generativnem modeliranju glasbe in ponuja bogate reprezentacije zvoka, vendar je zaradi svoje velikosti (5B parametrov) računsko zelo zahteven. M2D model [**?**] predstavlja učinkovitejšo alternativo, saj omogoča pridobivanje kakovostnih reprezentacij z bistveno manjšimi računskimi zahtevami.

III. Izbira Modela

A. Prvotna Zasnova z Jukeboxom

Projekt smo prvotno zasnovali z uporabo Jukebox modela zaradi njegovih naprednih zmožnosti reprezentacije glasbe:

- Bogat nabor značilk zaradi velikosti modela (5B parametrov)
- Predučenje na 1.2M skladbah različnih žanrov
- 4800-dimenzionalne reprezentacije za vsak segment

Vendar so se pri implementaciji pokazale pomembne omejitve:

- Potreba po vsaj 16GB GPU pomnilnika
- Čas inference več kot 30 sekund na posnetek
- Zahtevno procesiranje in kvantizacija zvoka
- Kompleksna postavitev modela v produkcijsko okolje

B. Prehod na M2D

Zaradi omenjenih omejitev smo sistem preoblikovali z uporabo M2D modela, ki ponuja:

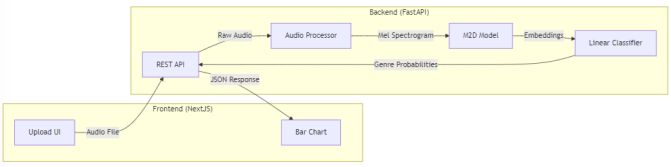
- Učinkovitejšo ViT-Base arhitekturo
- Inferenca v 7 sekundah na standardni GPU
- Potreba po samo 4GB GPU pomnilnika
- Enostavnejše procesiranje zvoka
- Primerljiva kvaliteta reprezentacij za klasifikacijo žanrov

IV. Metodologija

A. Arhitektura Sistema

Razviti sistem je sestavljen iz štirih glavnih komponent:

- Modul za procesiranje zvoka in pridobivanje M2D embeddingov
- Učno okolje za treniranje linearnega klasifikatorja
- FastAPI strežnik za serviranje modela
- React spletni vmesnik za interakcijo z uporabniki



Slika 1. Arhitektura sistema prikazuje tok podatkov od zvočnega posnetka do končne napovedi. Sistem je sestavljen iz štirih glavnih komponent: procesiranja zvoka, M2D modela, linearnega klasifikatorja in spletnega vmesnika.

B. Procesiranje Zvoka

Za zagotavljanje optimalnega delovanja M2D modela smo implementirali naslednje korake procesiranja:

- Pretvorba stereo posnetkov v mono z povprečenjem kanalov
- Prevzorčenje na 16000 Hz
- Segmentacija z oknom velikosti 25ms in preskokom 10ms
- Izračun 80 mel-pasovnih filtrov v razponu 50-8000 Hz
- Standardizacija z mean=-7.1 in std=4.2

C. M2D Model in Embeddingi

Osnova sistema je prednaučeni M2D model z naslednjo konfiguracijo:

- ViT-Base arhitektura
- Vhod dolžine 6 sekund
- Velikost zaplat (patch size) 16x16
- Izhodni embeddingi dimenzij [batch, time, 3840]
- Končne reprezentacije dobljene s povprečenjem po časovni dimenziji

D. Učenje Klasifikatorja

Za učenje linearnega klasifikatorja smo implementirali naslednji pristop:

- Zamrznjena M2D osnova ostane nespremenjena
- SGD optimizator z učno hitrostjo 0.1
- Implementacija zgodnjega ustavljanja
- Shranjevanje najboljšega modela glede na validacijsko množico
- Beleženje kontrolnih točk z vsemi potrebnimi stanji

V. Implementacija in Prispevki

A. Lastni Prispevki

V okviru projekta smo samostojno implementirali:

- Celoten cevovod za procesiranje zvoka v skladu z M2D zahtevami
- Sistem za učenje in evaluacijo linearnega klasifikatorja
- FastAPI strežnik z optimiziranim procesiranjem zahtev
- Interaktivni spletni vmesnik z vizualizacijo napovedi

B. Uporabljene Obstoječe Komponente

Projekt temelji na naslednjih obstoječih komponentah:

- Prednaučen M2D model korporacije NTT
- PyTorch ogrodje za implementacijo modela
- FastAPI za razvoj REST API-ja
- React in D3.js za spletni vmesnik

VI. Evaluacija in Rezultati

A. Časovna Zahtevnost

Analiza po komponentah:

- Procesiranje zvoka: O(n) za n vzorcev
- M2D embeddingi: O(n) z visokim konstantnim faktorjem
- Učenje klasifikatorja: O(k * m * d) za k epoh, m primerov, d dimenzij
- Inferenca: O(n) + O(d) za procesiranje in klasifikacijo

Izmerjeni časi izvajanja:

- Procesiranje 6s posnetka: 0.5s
- Izračun M2D embeddingov: 6s (GPU: GTX 1070 Ti)
- Učenje (100 epoh): 5 min
- Celotna inferenca: 7s

B. Prostorska Zahtevnost

Analiza porabe pomnilnika:

- M2D model: 300MB
- Medpomnenje spektrogramov: O(n)
- Embeddingi: 15KB na posnetek
- Linearni klasifikator: 30KB
- GPU pomnilnik med inferenco: 2GB

C. Sistemske Zahteve

Minimalne zahteve za delovanje:

- GPU: 4GB VRAM
- Disk: 500MB za model in knjižnice
- RAM: 2GB za inferenco
- Podprti formati: WAV, MP3, OGG
- Največja dolžina posnetka: 10 min

D. Primerjava z CNN Pristopom

Za ovrednotenje učinkovitosti našega pristopa smo implementirali tudi konvolucijsko nevronsko mrežo (CNN) s štirimi konvolucijskimi sloji, ki deluje neposredno na mel-spektrogramih. Primerjava rezultatov je pokazala:

Model	Točnost na testni množici
M2D + linearni klasifikator	76.0%
CNN model	61.5%

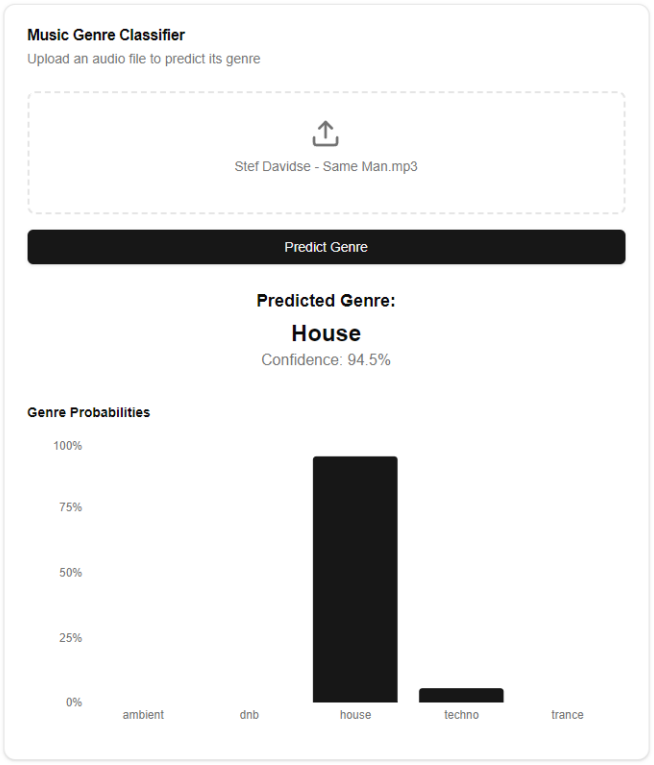
Tabela I
Primerjava točnosti različnih pristopov na testni množici.

Rezultati kažejo, da je pristop z uporabo prednaučenih M2D embeddingov bistveno uspešnejši od klasičnega CNN pristopa. To nakazuje, da M2D model uspešno zajame kompleksne značilnosti glasbenih posnetkov, ki jih je težje neposredno naučiti s konvolucijsko nevronsko mrežo. Višja točnost M2D pristopa je verjetno posledica predučenja na veliki količini glasbenih podatkov, kar omogoča boljšo generalizacijo tudi na manjši množici podatkov za specifičen žanr.

E. Uporabniški Vmesnik

Razviti spletni vmesnik ponuja:

- Preprost obrazec za nalaganje datotek
- Sprotno procesiranje in klasifikacijo
- Vizualizacijo zaupanja za vse žanre
- Odzivno oblikovanje za različne naprave



Slika 2. Spletni vmesnik omogoča nalaganje zvočnih datotek in prikazuje rezultate klasifikacije v obliki interaktivnega stolpičnega diagrama. Za vsak žanr je prikazana stopnja zaupanja v napoved.

VII. Zaključek

V članku smo predstavili celovito rešitev za klasifikacijo elektronske glasbe v podžanre. Začetna zasnova s kompleksnim Jukebox modelom je zaradi računskih omejitev vodila do izbire učinkovitejšega M2D modela, ki omogoča praktično uporabo na standardni strojni opremi ob ohranitvi visoke kakovosti klasifikacije. Ključni prispevki vključujejo razvoj optimiziranega cevovoda za procesiranje zvoka, implementacijo učinkovitega učnega sistema in razvoj uporabniku prijaznega vmesnika.

Primerjalna analiza s CNN modelom je potrdila ustreznost izbire M2D modela, saj je pristop z M2D reprezentacijami dosegel bistveno višjo točnost klasifikacije (76.0% proti 61.5%). To potrjuje vrednost uporabe prednaučenih modelov za zajem značilnosti glasbe, še posebej pri manjših podatkovnih množicah.

Možne izboljšave sistema vključujejo:

- Implementacijo paketnega procesiranja
- Kvantizacijo modela za zmanjšanje porabe pomnilnika
- Predpomnjenje embeddingov
- Paralelizacijo procesiranja

Literatura

[1] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2024). "Masked Modeling Duo: Towards a Universal Audio Pre-training Framework." *IEEE/ACM Trans. Audio, Speech, Language Process.*, 32, 2391-2406.

[2] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., Yasuda, M., Tsubaki, S., & Imoto, K. (2024). "M2D-CLAP: Masked Modeling Duo Meets CLAP for Learning General-purpose Audio-Language Representation." *to appear at Interspeech*.

[3] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2023). "Masked Modeling Duo: Learning Representations by Encouraging Both Networks to Model the Input." *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

[4] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2023). "Masked Modeling Duo for Speech: Specializing General-Purpose Audio Representation to Speech using Denoising Distillation." *Proc. INTERSPEECH 2023*, 1294-1298.

[5] Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2024). "Exploring Pre-trained General-purpose Audio Representations for Heart Murmur Detection." *to appear at IEEE EMBC*.

[6] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). "Jukebox: A generative model for music." *arXiv preprint arXiv:2005.00341*.

[7] Logan, B. (2000). "Mel Frequency Cepstral Coefficients for Music Modeling." *ISMIR 2000*.

[8] Choi, K., Fazekas, G., & Sandler, M. (2016). "Automatic tagging using deep convolutional neural networks." *ISMIR 2016*.

[9] Chen, Y.-P., et al. (2021). "MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training." *ACL-IJCNLP 2021*.