Prepoznavanje žanra muzike i emocije u pesmi

Anđela Trajković, sw76/2017 Milica Poparić, sw21/2017 Teodora Nedić, sw41/2017
Mentor: Milica Škipina
Fakultet tehničkih nauka, Univerzitet u Novom Sadu

Uvod

Cilj projekta je detekcija žanra i prepoznavanje emocije u pesmi na osnovu zvučnog isečka u trajanju od 30 sekundi. Projekat podržava detekciju 10 žanrova u koje spadaju bluz, klasična muzika, kantri, disko, hip hop, džez, metal, pop, rege i rok. Emocije koje smo detektovale u projektu: sreća, tuga, bes, relaksacija.
Sistem bi našao primenu u personalizovanju muzičkih predloga, na primer kreiranja plejliste na osnovu najslušajnijih žanrova kao i čestih raspoloženja korisnika.

Skup podataka

Podatke smo preuzele sa linka, izabrani skup podataka sadrži 30 sekundi audio zapisa u wav formatu za 10 žanrova pri čemu svaki žanr ima 100 pesama. Kako nam je skup podataka podeljen na 10 žanrova po folderima, bilo je potrebno označiti fajlove po emocijama. Fajlovi su ručno labelirani u csv fajlu našom subjektivnom procenom emocija u njima.

Metodologija

Problemu je pristupljeno na više načina, zarad utvrđivanja najboljih metodologija u rešavanju istog. Nad skupom podataka smo za sve mreže odradile podelu na trening, validacioni i test skup u razmeri 80:10:10.

Prvi pristup

- audio fajlovi su dekodirani pomoću tensorflow biblioteke,
- nad svim podacima primenjen algoritam *fast fourier transformations*
- za obučavanje je korišćena CNN mreža, sa
- *Adam* optimizacijom i
- *sparse categorical entropy* loss funkcijom

Drugi pristup

- interpretira audio zapise preko njihovih izdvojenih osobina, među kojima su i *MFCCs*
- nad izdvojenim podacima je urađena normalizacija, i
- kao klasifikator smo koristile *feed-forward* neuronsku mrežu,
- *Adam* optimizaciju i
- *sparse categorical entropy* loss funkciju

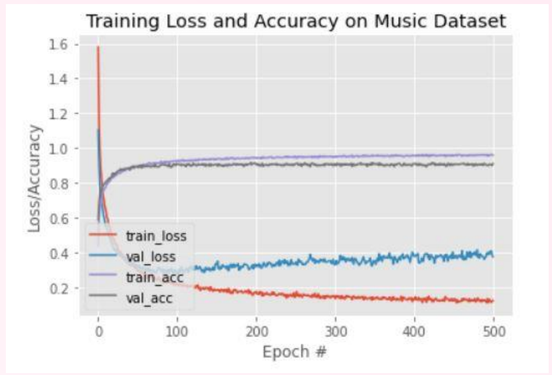
Ovaj pristup je pokazao veoma dobre rezultate u klasifikaciji žanra, kao i emocije uz minimalne promene neuronske mreže.

Rezultati

Prvi pristup sa FFT-om je dao odlične rezultate na trening i validacionom skupu, međutim test je imao samo 39% tačnosti, pa je ovaj način rada odbačen nakon nekoliko neuspešnih pokušaja da se prevaziđe overfitting.
U drugom pristupu, koji se pokazao kao mnogo bolji u prepoznavanju i žanrova i emocija, dobili smo sledeće rezultate:

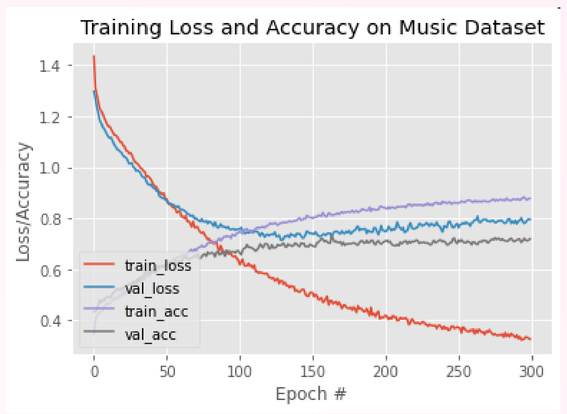
	train	valid	test
accuracy	0.95	0.91	0.90
loss	0.12	0.37	0.40

Tabela 1 Rezultati prepoznavanja žanrova



	train	valid	test
accuracy	0.87	0.71	0.71
loss	0.33	0.84	0.78

Tabela 2 Rezultati prepoznavanja emocija



Zaključak

FFT uzima audio signal i određuje "sadržaj frekvencije" signala. MFCC su čulno motivisane osobine zvuka koje se popudaraju sa načinom na koji ljudi doživljavaju visinu tona (ljudski sluh nije ravnomerno raspoređen, bolje razaznaje niže frekvencije od viših).
Možemo zaključiti da je model neuronske mreže koji koristi MFCC kao ulaz, bolje "naučio" da generalizuje na novim (nevidljivim) podacima.
Isti model je u prepoznavanju emocija postigao nešto lošije rezultate u odnosu na prepoznavanje žanrova. Pretpostavljamo da je mogući razlog labeliranje audio zapisa, koje je rađeno isključivo na osnovu našeg osećaja.
Povećanjem obima skupa podataka postoji mogućnost da bi se ostvarili bolji rezultati