Lumen - Data Science 2023

Projektna Dokumentacija

Ime tima: Tamburaši

Marko Haralović

Tin Josip Čurik

1 Uvod

Glazba je jedna od ključnih umjetnosti koja je prisutna u našim životima i igra važnu ulogu u našoj kulturi. S razvojem tehnologije, streaming servisi za glazbu nastoje poboljšati korisničko iskustvo i personalizirati preporuke sadržaja svojim korisnicima. Identifikacija instrumenata u pjesmama jedna je od najvažnijih informacija koje mogu biti dohvaćene od strane ovih servisa za poboljšanje preporuka i personalizacije sadržaja. Automatska klasifikacija instrumenata i vokala u pjesmama ima širok raspon primjena, uključujući lakše pretraživanje, točnije preporuke i daljnje detaljne i složene analize. Međutim, automatizacija ovog procesa nije trivijalna, s obzirom na visoku entropiju informacija u audio signalima, širok raspon izvora, postupak miksanja i teškoće analitičkog opisa. U ovom projektu ćemo se usredotočiti na istraživanje i implementaciju metoda za automatiziranu klasifikaciju instrumenata i vokala u pjesmama pomoću Dubokog Učenja, te analizirati učinkovitost tih metoda u praksi. Cilj ovog projekta je razviti i testirati dubokoučeni model za klasifikaciju instrumenata i vokala koji će postići visoku točnost klasifikacije na IRMAS-ovom skupu podataka. Ovaj rad će se također usredotočiti na analizu arhitekture mreže i različitih tehničkih aspekata učenja, kako bi se pružio uvid u učinkovitost različitih pristupa za klasifikaciju instrumenata i vokala u pjesmama. Konačno, ovaj rad će pokazati kako se primjena dubokog učenja može koristiti za automatiziranu klasifikaciju instrumenata i vokala, što ima potencijal za poboljšanje kvalitete usluga streaming servisa za glazbu i personalizacije sadržaja za korisnike.

2.1 Opis dataset-a

IRMAS dataset je skup glazbenih audio isječaka s anotacijama o istaknutim instrumentima uključenim u svakom isječku. Namijenjen je za automatsku identifikaciju istaknutih instrumenata u glazbi te je pogodan za razvoj i testiranje algoritama strojnog učenja. Dataset uključuje glazbu različitih desetljeća prošlog stoljeća, što pruža širok raspon kvalitete zvuka. Također pokriva raznolik raspon vrsta glazbenih instrumenata, tehnika sviranja, stilova snimanja i produkcije, te izvođača.

Podaci su podijeljeni u skupove za treniranje i testiranje, a svi audio datoteke su u 16-bitnom stereo formatu s brzinom uzorkovanja od 44.100 Hz. Skup za treniranje uključuje 6.705 audio datoteka s isječcima trajanja od 3 sekunde iz više od 2.000 različitih snimaka. Skup za testiranje sastoji se od 2.874 audio datoteka duljine od 5 do 20 sekundi, koje ne sadrže nijednu datoteku iz skupa za treniranje. Skup za testiranje sadrži jedan ili više ciljnih istaknutih instrumenata. Ukupni broj oznaka za treniranje jednak je broju audio datoteka, dok je broj oznaka za testiranje veći od broja audio datoteka za testiranje, jer su potonje višestruko označene. Važno je napomenuti da druge glazbene instrumente, poput perkusija, bubnjeva i basa, nisu uključene u anotaciju, čak i ako se pojavljuju u isječcima glazbe

.

Instruments	Abbreviations	Training (n)	Testing (n)
Cello	cel	388	111
Clarinet	cla	505	62
Flute	flu	451	163
Acoustic guitar	acg	637	535
Electric guitar	elg	760	942
Organ	org	682	361
Piano	pia	721	995
Saxophone	sax	626	326
Trumpet	tru	577	167
Violin	vio	580	211
Voice	voi	778	1044

U odluci o sastavljanju IRMAS dataseta, odlučili smo zadržati dijelove trening seta koji sadrže bubnjeve i druge šumove, umjesto da ih izbacimo iz skupa podataka. Ovo je učinjeno s ciljem zadržavanja robusnosti modela u prepoznavanju istaknutih instrumenata u uzorcima koji sadrže šum, uključujući bubnjeve. Očekujemo da će se slični šumovi pojaviti i u test setu, pa je bilo važno zadržati ove primjere u trening setu kako bi model bio sposoban prepoznati i klasificirati istaknute instrumente u takvim situacijama.

2.2 Data augmentation

Da bismo u IRMAS dataset uveli polifoniju, odlučili smo primijeniti augmentaciju podataka koja uključuje slučajno odabiranje pojedinačnih audio datoteka iz skupa za treniranje i njihovo spajanje u nove audio datoteke. U svakoj novoj datoteci smo slučajno odabrali između 1 i 5 istaknutih instrumenata koji sviraju u isto vrijeme. Na ovaj način smo stvorili novi skup podataka koji se sastoji od preko 20 000 isječaka, svaki po sekundu trajanja. Međutim, važno je napomenuti da takvi sintetički generirani podaci možda nisu adekvatna reprezentacija pravih polifonih traka, što predstavlja određeni rizik u korištenju ovakvog skupa podataka za obuku modela. Međutim, zbog nedostatka polifonih skupova podataka u glazbenoj klasifikaciji, ova tehnika je bila nužna kako bi se stvorio skup podataka koji se može koristiti za obuku modela koji je sposoban klasificirati polifoniju.

2.3 Procesiranje podataka

Prije nego što smo primijenili postupke obrade signala poput izrade spektrograma, kromagrama i spektralnih kontrasta, morali smo prvo obraditi sami zvukovni materijal. Budući da je IRMAS skup podataka u formatu wav, prvo smo ih morali dekompresirati i zatim isjeći u isječke duljine 1 sekunde kako bismo osigurali dosljedan ulaz u mrežu. Nakon toga smo primijenili proces downsampleiranja kako bismo smanjili uzorkovanje zvuka s izvornih 44.1kHz na 22.05kHz, što je dovoljno za analizu frekvencija u rasponu do 11kHz, čime smo uštedjeli na memorijskom prostoru. Zatim smo sve zvukove downmixali u mono kanal, što je u ovom slučaju bilo prikladnije jer IRMAS skup podataka nije snimljen u surround formatu. Nakon toga smo primijenili normalizaciju Root Mean Square Energyom (RMS), kojom smo osigurali da svi zvukovi imaju sličnu glasnoću, što je važno kako bi se izbjegle neželjene fluktuacije glasnoće tijekom treniranja i testiranja modela. Tek nakon ovih koraka smo mogli primijeniti postupke obrade signala koji slijede.

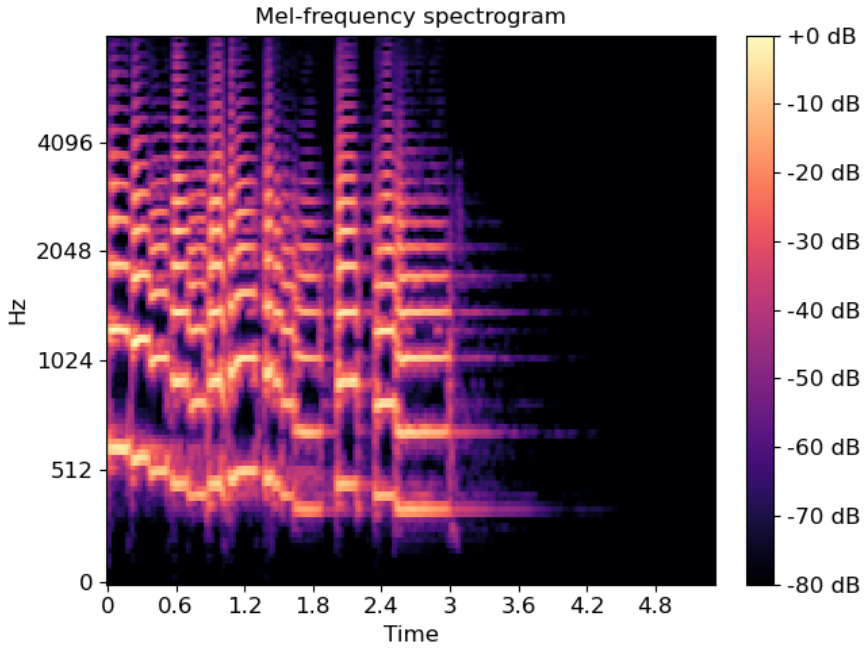
2.3.2 Mel Spektrogrami

Mel-spektrogrami su popularna reprezentacija zvučnog signala koja se koristi u području obrade zvuka, posebno u glazbenoj klasifikaciji. Mel-spektrogrami su grafički prikazi frekvencijskog sadržaja zvuka u ovisnosti o vremenu, stvoreni pomoću algoritama za transformaciju Fourierovih spektara zvuka.

Proces stvaranja Mel-spektrograma uključuje nekoliko koraka. Prvo, audio datoteka u formatu .wav se uzorkuje na određenoj brzini uzorkovanja, što se obično odabire na 44,1 kHz. Zatim, primjenjuje se kratkotrajna Fourierova transformacija (engl. Short-time Fourier transform, STFT) na audio datoteku kako bi se dobila informacija o frekvencijskom sadržaju zvuka u ovisnosti o vremenu. STFT se primjenjuje na mali segment audio signala, obično oko 20 do 50 milisekundi duljine, s preklapanjem između segmenata. Rezultat primjene STFT-a je spektrogram koji prikazuje amplitudu svake frekvencije zvuka u odnosu na vrijeme.

Nakon dobivanja spektrograma, primjenjuje se proces Mel-frekvencijske skaliranja (engl. Mel-frequency scaling) kako bi se pretvorio spektrogram u Mel-spektrogram. Mel-frekvencijsko skaliranje se koristi za transformaciju linearno razmještenih frekvencija u Mel-skalu koja je osjetljivija na male promjene u nižim frekvencijama. Mel-skala se temelji na ljudskom sluhu, što je čini prikladnom za analizu zvuka u glazbenoj klasifikaciji.

Nakon što se spektrogram pretvori u Mel-spektrogram, mogu se primijeniti razni dodatni postupci obrade signala, poput normalizacije, filtriranja šumova i komprimiranja, kako bi se pripremili za korištenje u procesu obuke modela strojnog učenja. Mel-spektrogrami su popularni izbor za reprezentaciju zvuka u glazbenoj klasifikaciji zbog svoje sposobnosti da prikazuju frekvencijski sadržaj zvuka u ovisnosti o vremenu i lakoće korištenja u modelima strojnog učenja.



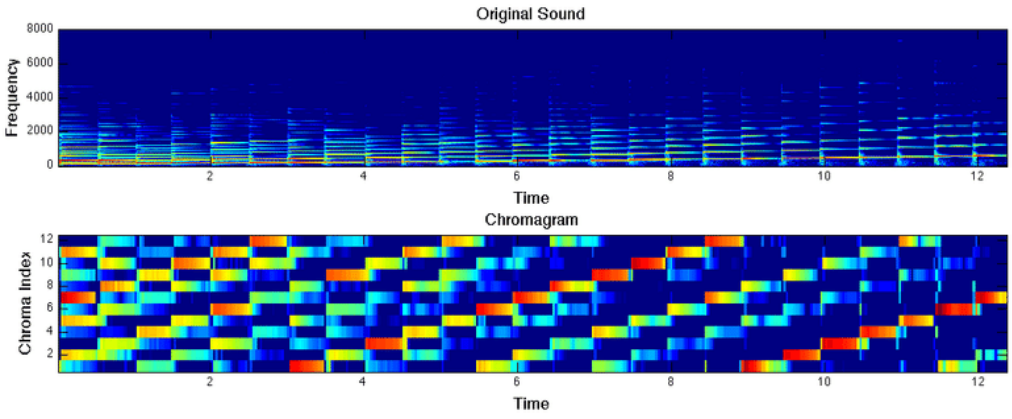
2.3.3 Kromagrami

Kromagrami (eng. chromagrams) su još jedna popularna reprezentacija zvučnog signala u glazbenoj klasifikaciji. Kromagrami se koriste za prikazivanje informacija o frekvencijama tonova u glazbi, što ih čini korisnim za analizu harmonijskih i melodskih obrazaca u glazbi.

Proces stvaranja kromagrama uključuje transformaciju Mel-spektrograma kako bi se dobile informacije o tonalitetu glazbe. Nakon dobivanja Mel-spektrograma, primjenjuje se postupak koji se naziva tonalitetni filtrirajući bank (eng. Tonal Filtering Bank). Tonalitetni filtrirajući bank sastoji se od 12 filtra, koji su postavljeni na različite frekvencijske pojasove, koji se odnose na 12 različitih tonaliteta u glazbi. Svaki filtar mjeri količinu energije u spektru zvuka u određenom tonalitetu. Kada se primijeni tonalitetni filtrirajući bank na Mel-spektrogram, dobivaju se 12 različitih izlaza koji odgovaraju intenzitetu zvuka za svaki tonalitet.

Nakon dobivanja izlaza tonalitetnog filtrirajućeg banka, primjenjuje se postupak normalizacije kako bi se kromagrami skalirali u raspon od 0 do 1. Kromagrami se mogu prikazati kao matrice, gdje svaki stupac predstavlja jedan oktavni pojas, a svaki red predstavlja jedan tonalitet. Kromagrami su vrlo korisni u glazbenoj klasifikaciji, jer omogućuju analizu harmonijske strukture i melodije glazbe. Oni se mogu koristiti za analizu akorda, harmonijske progresije, izmjene tonaliteta i druge karakteristike glazbenih kompozicija.

Ukratko, kromagrami su prikazi frekvencijskog sadržaja zvuka u ovisnosti o tonalitetu i oktavnom rasponu. Kromagrami su korisni za analizu harmonijskih i melodskih obrazaca u glazbi, što ih čini korisnim alatom za razvoj modela strojnog učenja u glazbenoj klasifikaciji.



2.3.4 Spektralni Kontrasti

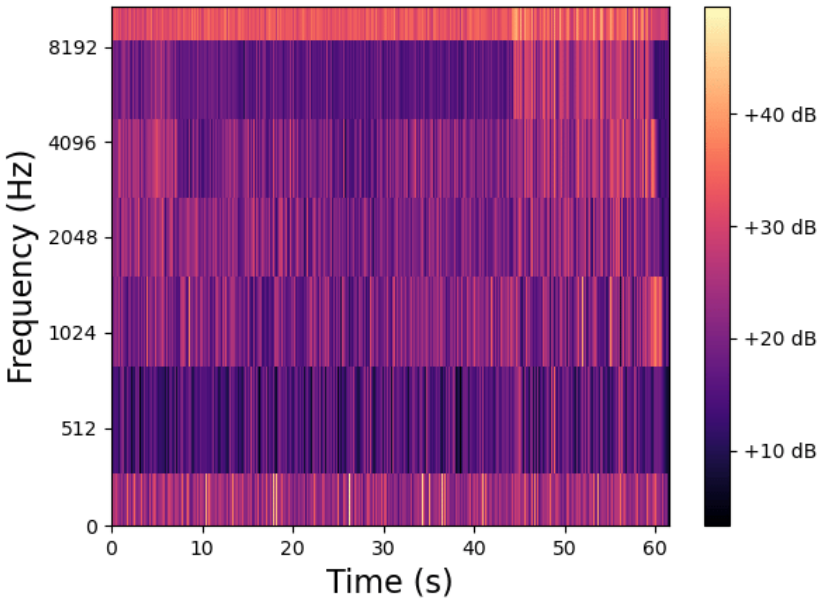
Spektralni kontrasti (eng. spectral contrasts) su još jedna popularna reprezentacija zvučnog signala u glazbenoj klasifikaciji. Spektralni kontrasti se koriste za prikazivanje razlike u intenzitetu zvuka na različitim frekvencijama, što ih čini korisnim za analizu harmonijskih i ritmičkih obrazaca u glazbi.

Proces stvaranja spektralnih kontrasta uključuje prvo dobivanje Mel-spektrograma. Nakon toga, primjenjuje se postupak koji se naziva računanje spektralne ravnoteže (eng. spectral balance). Spektralna ravnoteža se računa tako da se podijeli spektrogram na 9 jednakih frekvencijskih pojasova, a zatim se izračuna srednja vrijednost intenziteta u svakom pojasu. Nakon dobivanja spektralne ravnoteže, izračunavaju se spektralni kontrasti za svaki oktavni pojas i za svaku kombinaciju frekvencijskih pojasova.

Spektralni kontrasti se mogu prikazati kao matrice, gdje svaki stupac predstavlja jedan oktavni pojas, a svaki red predstavlja kontrast između dva frekvencijska pojas. Veći kontrast ukazuje na veću razliku u intenzitetu zvuka između dva frekvencijska pojas.

Spektralni kontrasti su korisni u glazbenoj klasifikaciji, jer omogućuju analizu harmonijske i ritmičke strukture glazbe. Oni se mogu koristiti za analizu dinamičke raznolikosti u glazbenim kompozicijama, što može biti korisno u prepoznavanju različitih glazbenih žanrova i emocionalnih sadržaja u glazbi. Spektralni kontrasti također mogu biti korisni u kombinaciji s drugim reprezentacijama zvuka, poput Mel-spektrograma i kromagrama, kako bi se poboljšala kvaliteta glazbene klasifikacije.

Ukratko, spektralni kontrasti su prikazi razlike u intenzitetu zvuka na različitim frekvencijama. Oni su korisni za analizu harmonijske i ritmičke strukture glazbe i mogu se koristiti za prepoznavanje različitih glazbenih žanrova i emocionalnih sadržaja u glazbi. Spektralni kontrasti su još jedan koristan alat u razvoju modela strojnog učenja u glazbenoj klasifikaciji.



2.3.5 Kombinacija navedenih reprezentacija

Mel-spektrogrami, kromagrami i spektralni kontrasti su tri različite metode reprezentacije zvuka koje se često koriste u glazbenoj klasifikaciji i prepoznavanju glazbenih uzoraka. Svaka od ovih metoda pruža jedinstvenu perspektivu na glazbu i različitu vrstu informacija.

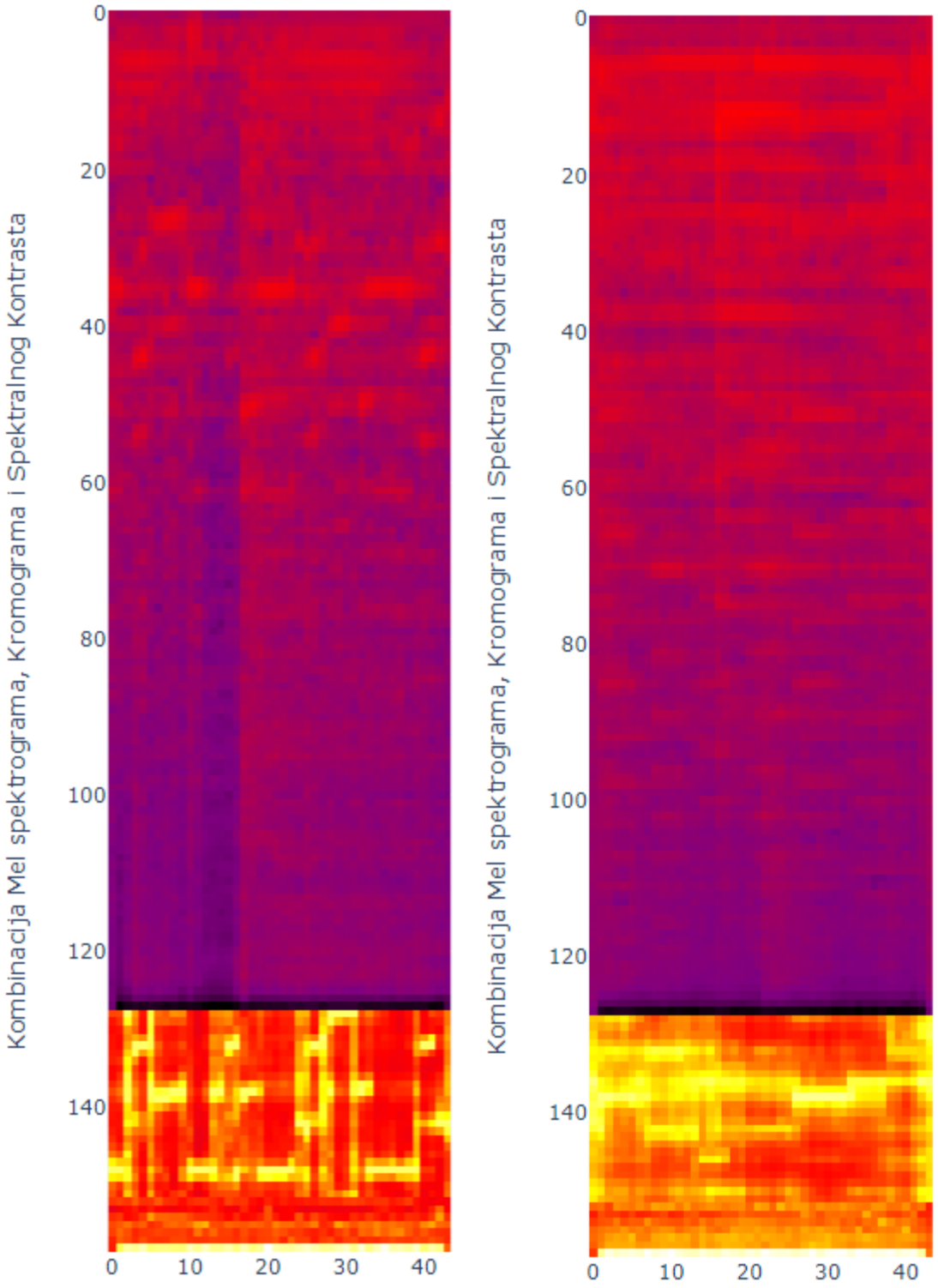
Mel-spektrogram pruža uvid u spektralnu gustoću zvuka, odnosno prikazuje jačinu različitih frekvencija tijekom vremena. Ova metoda je korisna za prepoznavanje ritma, kao i za identifikaciju pojedinačnih instrumenata i vokala u glazbi.

Kromagram, s druge strane, prikazuje raspored tonova i tonalitet pjesme. Ova metoda se koristi za prepoznavanje harmonije i tonaliteta, kao i za razlikovanje različitih vrsta glazbenih žanrova.

Spektralni kontrasti predstavljaju razlike između spektara različitih vremenskih segmenata. Ova metoda može biti korisna za prepoznavanje dinamičkih uzoraka u glazbi, poput staccata, legata ili glasnoće. Također se može koristiti za prepoznavanje različitih glazbenih žanrova.

Kombiniranjem ovih triju metoda, može se dobiti sveobuhvatniji prikaz zvuka i pružiti više informacija o glazbi koja se analizira. To može pomoći u poboljšanju preciznosti prepoznavanja glazbenih uzoraka, posebno kada se radi o složenim glazbenim žanrovima koji imaju različite ritmove, harmonije i dinamičke uzorke.

Iako kombiniranje tih triju reprezentacija zahtijeva nešto više računalne snage i vremena, to pruža značajno poboljšanje u točnosti prepoznavanja glazbenih uzoraka. Stoga je ova kombinacija vrlo korisna u modeliranju strojnog učenja za klasifikaciju glazbe.


Kombinacija Mel spektrograma, Kromograma i Spektralnog Kontrasta
0
20
40
60
80
100
120
140
0
10
20
30
40
Kombinacija Mel spektrograma, Kromograma i Spektralnog Kontrasta
0
20
40
60
80
100
120
140
0
10
20
30
40

2.3.6 Pristup problemu

Mel-spektrogrami, spektralni kontrasti i kromagrami su, u biti, prikazi audio signala u 2D obliku, slično kao što su slike prikazi vizualnih podataka u 2D obliku. U audio procesiranju, ovakvi prikazi se koriste kako bi se audio signal preveo u oblik koji je lakše obraditi i analizirati. Konkretno, u našem slučaju, ovi prikazi su korišteni za pretvaranje problema prepoznavanja instrumenata u glazbi u problem detekcije slika, što nam omogućuje primjenu tehnika strojnog učenja koje se koriste za obradu slika.

Da bismo primijenili konvolucijske neuronske mreže za ovaj problem, morali smo prikaze audio signala transformirati u oblik koji se može koristiti kao ulazni podatak u mrežu. To smo postigli tako da smo prikaze pretvorili u slike, pri čemu svaki piksel predstavlja određenu vrijednost u tom prikazu. Tada smo primijenili tehnike obrade slika poput konvolucije i poolinga kako bismo izgradili konvolucijsku neuronsku mrežu koja može naučiti prepoznavati različite instrumetne uzorke u glazbi.

Ovakav pristup se pokazao vrlo uspješnim u području glazbene klasifikacije i prepoznavanja instrumenata, i trenutno je state-of-the-art pristup za ovaj problem. Konvolucijske neuronske mreže se mogu trenirati na velikim skupovima podataka, što omogućuje stvaranje modela koji je sposoban prepoznati različite uzorke i varijacije u glazbi, te može biti vrlo koristan alat u različitim glazbenim aplikacijama.

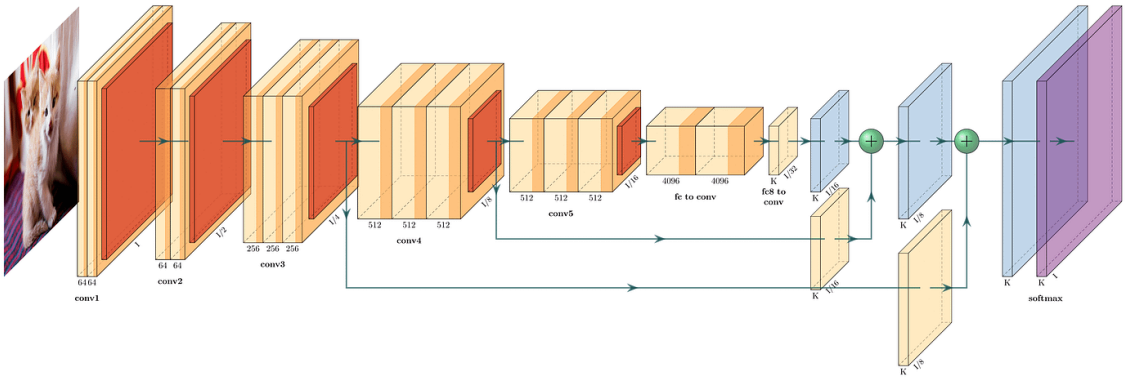
3. Konvolucijske neuronske mreže

Konvolucijske neuronske mreže (CNN) su vrsta neuronskih mreža koja su posebno učinkovite u rješavanju problema prepoznavanja uzoraka u slikama. Ova vrsta neuronskih mreža se sastoji od slojeva koji obrađuju ulazne podatke konvolucijskim filtrima koji su optimizirani za detekciju specifičnih oblika i značajki. Nakon toga, ulazni podaci se reduciraju i obrađuju kroz slojeve koji izvode operacije kao što su pooling i normalizacija. Konačno, ulazni podaci se pretvaraju u vektor značajki koji se prosljeđuje kroz sloj potpuno povezanih neurona koji izvršavaju klasifikaciju.

CNN se koriste u brojnim aplikacijama, poput prepoznavanja lica, detekcije objekata i prepoznavanja rukom pisanih slova. Kada se primjenjuju na audio podatke, ulazni podaci su predstavljeni u obliku 2D spektrograma koji prikazuju spektralni sadržaj zvuka u ovisnosti o vremenu. Ova transformacija audio podataka u 2D oblik omogućuje primjenu CNN za obradu audio podataka, pretvarajući problem klasifikacije zvuka u problem klasifikacije slika.

Mel-spektrogrami, spektralni kontrasti i kromagrami su glavni načini za predstavljanje audio podataka u 2D obliku, što omogućuje primjenu CNN-a na audio podatke. Te metode audio procesiranja pretvaraju audio podatke u spektrogram, koji je vizualizacija zvuka u ovisnosti o frekvenciji i vremenu. Nakon pretvorbe u spektrogram, primjenjuju se dodatni koraci za stvaranje kromagrama ili spektralnih kontrasta.

Korištenje CNN-a za klasifikaciju audio podataka postalo je state-of-the-art pristup zbog svoje sposobnosti da automatski nauče značajke iz podataka i stvore diskriminatorni model. Korištenje CNN-a za klasifikaciju audio podataka također omogućuje brzu i učinkovitu obradu velike količine podataka, što je posebno važno u glazbenoj klasifikaciji gdje postoji velika količina zvučnih zapisa.



3.1 Arhitektura neuralne mreže

4x	2x Conv2D (3 x 3)
	Batch Normalization
	LeakyReLU (a = 0.3)
	Max Pooling (p)
	Dropout (0.2)
	Dense (1024)
	LeakyReLU (a = 0.3)
	Batch Normalization
	Dropout (0.5)
	Dense (11)
	Sigmoid Activation

Odabir strukture naše mreže temeljio se na pregledu postojećih radova u području klasifikacije instrumenata i audio klasifikacije općenito, koji su pokazali da su konvolucijske neuronske mreže vrlo učinkovite u ovim zadacima. Iz tog pregleda, izdvojili smo nekoliko radova koji su se istaknuli po svojoj uspješnosti, a zajednička im je bila upotreba sličnih struktura mreže. Tako smo odlučili koristiti konvolucijske slojeve koji su praćeni slojevima BatchNormalization i aktivacijskim funkcijama LeakyReLU, kao i slojevima MaxPooling i Dropout za smanjenje dimenzionalnosti i regulaciju mreže.

Nakon što smo postavili osnovnu strukturu mreže, eksperimentirali smo s različitim arhitekturama kako bismo postigli optimalne performanse. Iz tog procesa eksperimentiranja, odabrali smo ovu strukturu koja se pokazala najučinkovitijom na našem skupu podataka. Struktura mreže ima tri para konvolucijskih slojeva s postupnim povećavanjem broja filtrirajućih jedinica. Nakon toga, slijede slojevi MaxPooling i Dropout koji smanjuju dimenzionalnost i reguliraju mrežu. Nakon toga, slijedi sloj Flatten koji priprema podatke za prolazak kroz potpuno povezani sloj s 1024 neurona i aktivacijskom funkcijom LeakyReLU. Nakon toga, slijedi sloj Dropout koji dodatno regulira mrežu te završni potpuno povezani sloj s 11 neurona i aktivacijskom funkcijom softmax.

Ukupan broj parametara mreže iznosi 9,628,811, a od toga je većina parametara povezana s potpuno povezanim slojevima. U konačnici, ovakva struktura mreže je pokazala vrlo dobre performanse u klasifikaciji instrumenata.

3.1.1 Konvolucijski slojevi

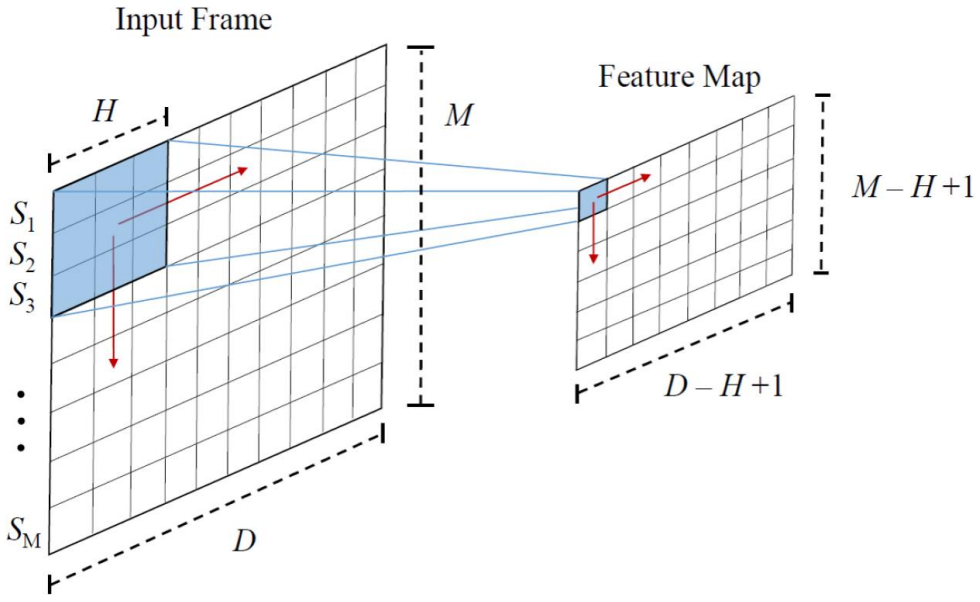
Konvolucijski slojevi su ključni elementi u konvolucijskim neuronskim mrežama (CNN) i imaju ključnu ulogu u obradi slike, prepoznavanju uzoraka i klasifikaciji objekata. Konvolucijski slojevi rade tako da primjenjuju filtere na ulaznu sliku ili matricu značajki, što rezultira izlaznom matricom značajki koja se koristi kao ulaz u sljedeći sloj.

Svaki konvolucijski sloj ima skup filtera, koji se često nazivaju jezgrama, a ti filteri su male matrice dimenzija (obično 3x3 ili 5x5) koje se primjenjuju na ulaznu sliku kako bi se stvorila nova matrica značajki. Kada se primijeni svaki filter na ulaznu sliku, generira se nova matrica značajki koja se koristi kao ulaz u sljedeći sloj. Važno je napomenuti da se u svakom konvolucijskom sloju primjenjuje više filtera, što omogućuje da se istovremeno otkrivaju različiti uzorci u ulaznoj slici.

Jedna od ključnih značajki konvolucijskih slojeva je da se svaki filter u sloju uči samostalno. Tijekom treninga, CNN će naučiti optimalne vrijednosti filtera za zadatak koji rješava. Ova sposobnost automatskog učenja filtera omogućuje konvolucijskim slojevima da otkrivaju složene uzorke u ulaznoj slici koji bi bili teško otkriti ručno.

Osim filtera, konvolucijski slojevi često koriste i operacije aktivacije poput ReLU (rectified linear unit) aktivacije, koja se koristi za uklanjanje negativnih vrijednosti iz matrice značajki i povećanje propusnosti signala. Ova aktivacija može poboljšati performanse modela i ubrzati konvergenciju tijekom treninga.

Konvolucijski slojevi su također osjetljivi na lokalne uzorke u ulaznoj slici, što omogućuje da se modeli fokusiraju na lokalne značajke koje su relevantne za klasifikaciju objekata. Ova sposobnost čini konvolucijske neuronske mreže vrlo učinkovitima za prepoznavanje uzoraka u slikama, kao što su prepoznavanje lica, prepoznavanje objekata u prometu i prepoznavanje rukopisa.



3.1.2 Batch Normalization

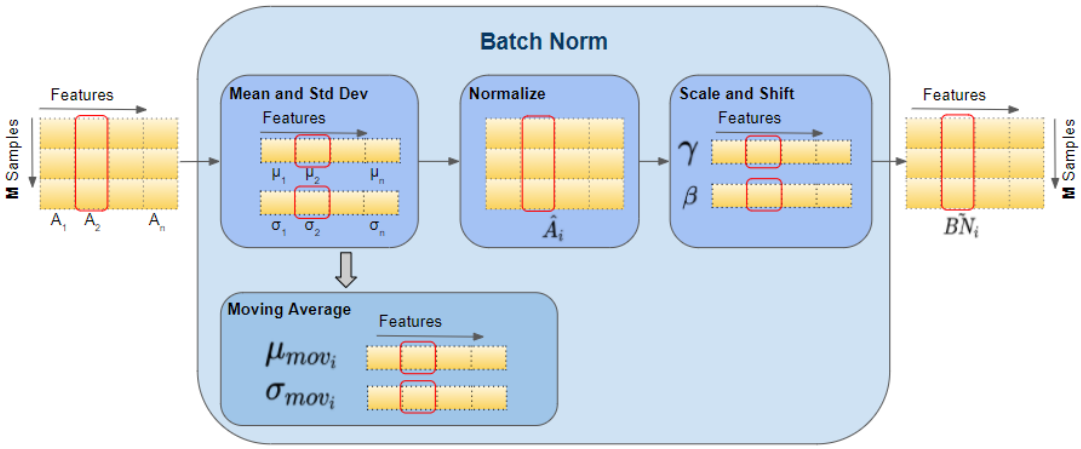
Batch Normalization (BN) je tehnika normalizacije podataka koja se koristi u neuronskim mrežama kako bi se poboljšala njihova stabilnost i učinkovitost u učenju. Ovaj postupak se primjenjuje na izlaz svakog sloja neuronske mreže, a cilj mu je normalizirati vrijednosti aktivacija koje izlaze iz sloja. Normalizacija se vrši na mini-batchevima, što znači da se statistike normalizacije računaju samo za taj mini-batch.

BN postupak se sastoji od dva glavna koraka. Prvi korak je centriranje i skaliranje podataka tako da se srednja vrijednost svakog feature-a postavi na 0, a varijanca na 1. Ovo se postiže stohastičkim gradijentnim spustom kako bi se naučili parametri skaliranja i pomaka koji se primjenjuju na svaki feature.

Drugi korak uključuje primjenu linearnog transformatora na normalizirane vrijednosti, a zatim primjenu nelinearne aktivacijske funkcije na izlazu transformatora. Ovaj postupak osigurava da vrijednosti aktivacija ostanu u prikladnom rasponu, čime se poboljšava sposobnost neuronskih mreža da nauče složene funkcije.

U neuronskim mrežama, BN se često koristi nakon konvolucijskih slojeva i prije aktivacijskih slojeva. To je zato što su konvolucijski slojevi skloni generiranju visokih vrijednosti, što može uzrokovati probleme u učenju kada se primjenjuju na sljedeći sloj. BN rješava taj problem normalizirajući izlaze iz slojeva i time sprječava probleme s eksplozijom gradijenta.

Korištenje BN-a ima nekoliko prednosti. Omogućuje neuronskim mrežama da rade s većim stopama učenja, što dovodi do bržeg učenja. Također sprječava prenaučenost (eng. overfitting) i poboljšava performanse mreže na novim primjerima. Osim toga, omogućuje korištenje neuronskih mreža s dubljim arhitekturama, što je postalo važno u raznim primjenama strojnog učenja. BN se također često koristi u kombinaciji s drugim tehnikama, poput dropout-a, što dodatno poboljšava stabilnost i učinkovitost neuronskih mreža.



3.1.3 Aktivacijske funkcije

Aktivacijske funkcije su ključne za rad neuronskih mreža jer određuju način na koji se neuroni aktiviraju i prijenose informaciju. Postoje razne aktivacijske funkcije koje se koriste u neuronskim mrežama, poput sigmoidne, tangens hiperbolne, ReLU (Rectified Linear Unit), Leaky ReLU, ELU (Exponential Linear Unit), itd.

ReLU funkcija je popularna u neuronskim mrežama zbog njene jednostavnosti i brzine izvođenja. Ona vraća vrijednost 0 za sve negativne ulaze, dok se pozitivni ulazi propuštaju kroz funkciju bez promjene. Međutim, ReLU funkcija ima problem "mrtvih neurona", gdje se neuroni aktiviraju samo za pozitivne vrijednosti, a za negativne ulaze vraćaju vrijednost 0. To znači da se težine tih neurona ne mijenjaju tijekom treniranja, što može dovesti do problema s učenjem.

Leaky ReLU funkcija je modifikacija ReLU funkcije koja rješava problem mrtvih neurona. Umjesto da vraća vrijednost 0 za negativne ulaze, ona vraća vrijednost koja je pomaknuta od nule za malu konstantu, čime se omogućuje korekcija težina i poboljšava se sposobnost neuronske mreže da nauči značajke. Leaky ReLU funkcija s parametrom alpha=0.33 se pokazala kao posebno prikladna za ovaj problem, a koristi se u našoj neuronskoj mreži za klasifikaciju instrumenata.

ELU funkcija je također popularna alternativa ReLU funkciji. Ona vraća vrijednost 0 za negativne ulaze, ali za pozitivne ulaze vraća eksponencijalnu vrijednost koja se pomaknuta za konstantu, čime se izbjegava problem mrtvih neurona i omogućava bolje učenje. Međutim, ELU funkcija je računski zahtjevnija od Leaky ReLU funkcije.

Istina je da smo odabrali Leaky ReLU funkciju sa parametrom alfa = 0.33 jer se u literaturi pokazala kao jedna od najefikasnijih funkcija za audio i instrument klasifikaciju. Leaky ReLU funkcija je jednostavna funkcija aktivacije koja se koristi u neuronskim mrežama. Ona djeluje tako da, za sve negativne vrijednosti ulaza, umjesto da ih postavi na nulu kao što to radi ReLU funkcija, pomnoži ih s parametrom alfa. Na ovaj način, Leaky ReLU funkcija omogućuje prolazak neznatno negativnih vrijednosti, što se pokazalo korisnim u klasifikacijskim zadacima.

Leaky ReLU funkcija ima nekoliko prednosti u odnosu na druge funkcije aktivacije, kao što su sigmoid i tanh. Jedna od njih je da Leaky ReLU funkcija ne pati od problema "mrtvih neurona", odnosno neurona koji ne aktiviraju uopće, što je čest problem kod ReLU funkcije. Leaky ReLU također ima manje oscilacija u odnosu na druge funkcije aktivacije, što je korisno u stabilizaciji učenja u neuronskim mrežama.

Activation func.	Micro			Macro		
	P	R	$F1$	P	R	$F1$
tanh	0.416	0.625	0.499	0.348	0.537	0.399
ReLU	0.640	0.550	0.591	0.521	0.508	0.486
PReLU	0.612	0.565	0.588	0.502	0.516	0.490
LReLU (α=0.01)	0.640	0.552	0.593	0.530	0.507	0.492
LReLU (α=0.33)	0.655	0.557	0.602	0.541	0.508	0.503

3.1.4 Max Pooling

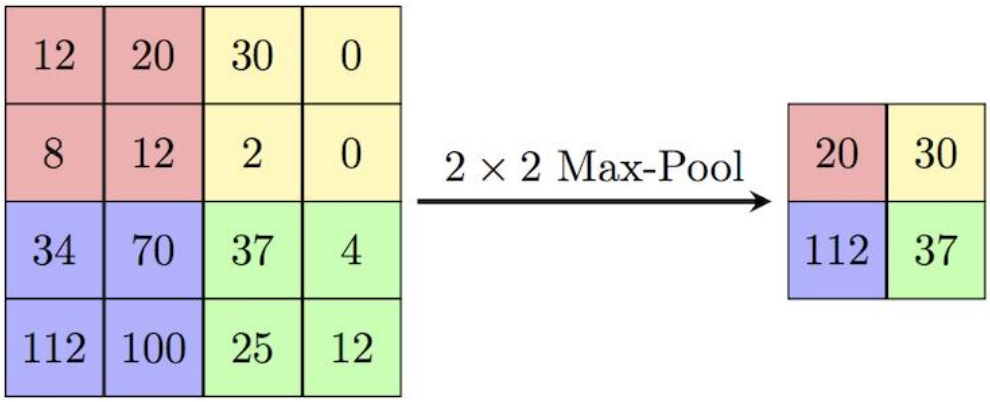
Max pooling je tehnika u računalnom vidu koja se često koristi u konvolucijskim neuronskim mrežama (CNN). Ova tehnika se koristi za smanjenje prostorne dimenzionalnosti izlaza konvolucijskog sloja, a time i smanjenje broja parametara u mreži.

Max pooling radi tako da za svaki podprostor (eng. receptive field) izlaza konvolucijskog sloja bira maksimalnu vrijednost. Ovaj postupak se obično radi na podprostorima veličine 2x2, pri čemu se koračanje vrši za 2 uzastopna elementa u svakoj dimenziji.

Ovaj proces ima nekoliko važnih funkcija u CNN-ovima. Prvo, smanjuje dimenzionalnost izlaza, što znači da se postiže brže izvršavanje, jer se smanjuje broj parametara koje mreža mora obraditi u kasnijim slojevima. Također, smanjenje dimenzionalnosti smanjuje količinu prostornih informacija koje se prenose u naredne slojeve, čime se postiže ujednačavanje i održavanje uzoraka bez obzira na poziciju u slici.

Međutim, postoji i nekoliko nedostataka max pooling-a. Prvo, postupak odabira maksimalne vrijednosti znači da se gube informacije o ostalim vrijednostima u podprostoru, što može dovesti do gubitka informacija o poziciji i obliku objekata u slici. Također, postupak koračanja za 2 elementa u svakoj dimenziji znači da se izgubi mogućnost detekcije objekata koji su manji od veličine podprostora koji se koristi za max pooling.

Uprkos ovim nedostacima, max pooling se i dalje često koristi u konvolucijskim neuronskim mrežama, jer se pokazao vrlo učinkovitim u postizanju dobrih performansi u mnogim problemima računalnog vida. U kontekstu IRMAS dataset-a, korištenje max pooling-a u mreži je doprinijelo bržoj konvergenciji i boljim rezultatima klasifikacije, s obzirom na količinu podataka koju smo imali.



3.2.5 Dropout

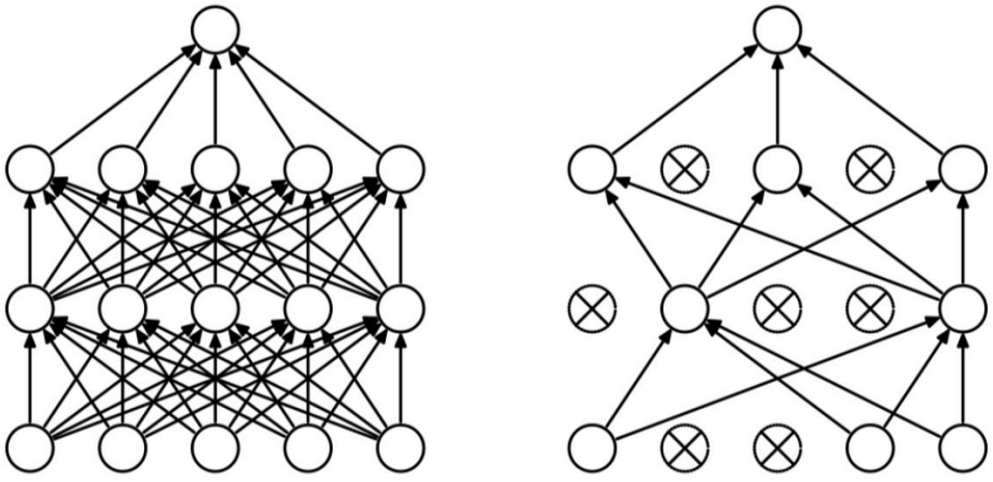
Dropout je tehnika regulacije koja se koristi u neuronskim mrežama kako bi se spriječilo pretjerano prilagođavanje (overfitting) modela podacima za učenje. Overfitting se događa kada model prekomjerno prilagođava svoje parametre podacima za učenje, što može dovesti do loših performansi na novim, neviđenim podacima.

Dropout radi tako da se, tijekom faze učenja, neke jedinice u mreži slučajno isključuju s vjerojatnošću p. To znači da se u svakoj epohi učenja neke jedinice u mreži ne koriste, što rezultira smanjenjem kapaciteta mreže i poticanjem neuronskih puteva da se prilagode različitim značajkama.

Na primjer, ako je p postavljen na 0,5, tijekom svakog prolaska kroz mrežu, polovica jedinica će se isključiti, a polovica će biti aktivne. Ovaj postupak se ponavlja tijekom svakog prolaska kroz mrežu.

Dropout ima nekoliko prednosti. Prvo, on sprječava pretjerano prilagođavanje i povećava sposobnost generalizacije modela. Drugo, on sprečava korelaciju između parametara u mreži, što može poboljšati učinkovitost učenja i brže konvergiranje. Treće, on može pomoći u borbi protiv problema s nebalansiranim klasama u skupu za učenje.

U konvolucijskim neuronskim mrežama, dropout se obično primjenjuje nakon maksimalne agregacije (max pooling). Ova tehnika se također može primijeniti na potpuno povezane slojeve (fully connected layers) u mreži. Međutim, važno je odabrati pravu vrijednost p kako bi se postigao optimalni učinak i izbjeglo prekomjerno izostavljanje jedinica, što može dovesti do gubitka važnih značajki i pogoršati performanse modela.



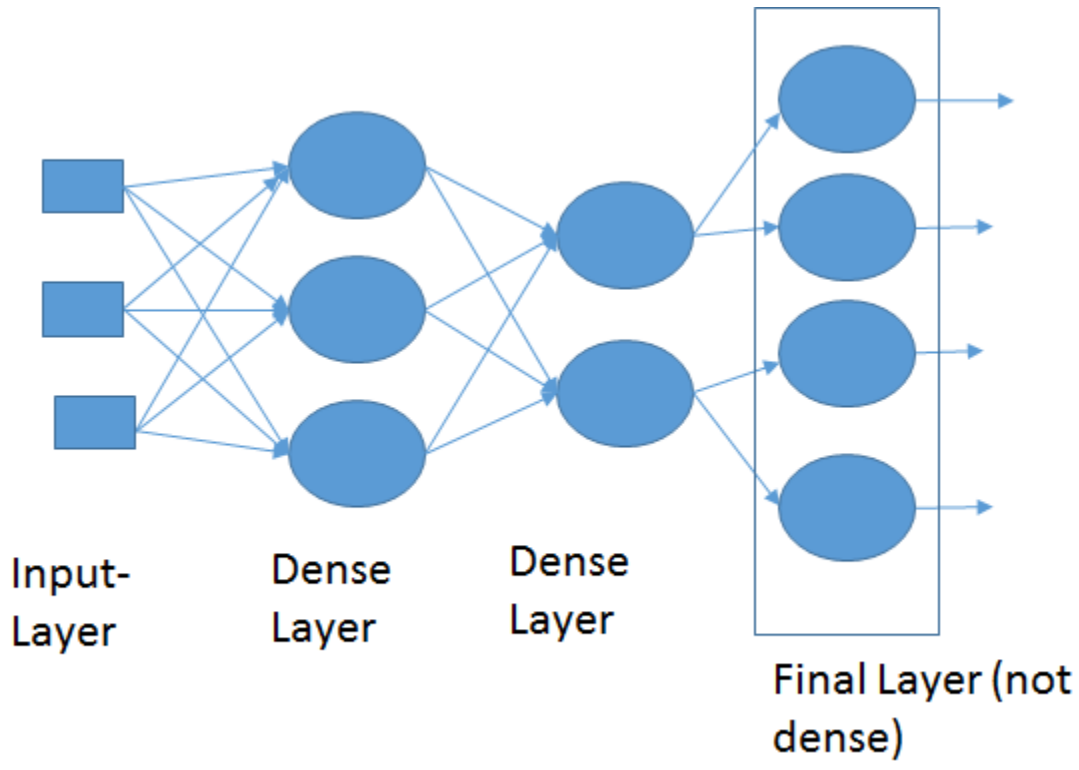
(a) Standard Neural Net (b) After applying dropout.

3.2.6 Dense Layer

Dense slojevi u konvolucijskim neuronskim mrežama obično slijede nakon konvolucijskih slojeva i služe za klasifikaciju. U slučaju IRMAS dataset-a, na kraju CNN arhitekture imamo zadnji dense sloj koji se sastoji od 11 neurona, jedan za svaku klasu koju želimo klasificirati. Prije toga, imamo drugi dense sloj s 1024 neurona koji služi za ekstrakciju značajki iz podataka koje su naučene u prethodnim slojevima i koje su najrelevantnije za klasifikaciju. U ovom sloju se također primjenjuje aktivacijska funkcija i batch normalization. Konačni izlaz iz ovog sloja predstavlja vektor značajki koji se prosljeđuje u zadnji dense sloj za klasifikaciju.

Broj neurona u dense slojevima ovisi o složenosti zadatka i broju klasa koje želimo klasificirati. U slučaju IRMAS dataset-a, odlučili smo se za 1024 neurona u drugom dense sloju zbog relativno velike složenosti zadatka klasifikacije instrumenata u različitim okolinama.

Kao i u drugim slojevima, u dense slojevima se primjenjuju aktivacijske funkcije i batch normalization kako bi se postigla veća generalizacija modela i poboljšala performansa. U ovom slučaju, aktivacijska funkcija koju smo koristili u dense slojevima bila je Leaky ReLU s alpha = 0.33, koja se pokazala kao najprikladnija u literaturi za ovakve vrste problema klasifikacije.



4. Zaključak

U ovom radu smo predstavili arhitekturu duboke konvolucijske neuronske mreže za klasifikaciju zvukova instrumenata. Uzeli smo IRMAS dataset koji se sastoji od raznih zvukova instrumenata te smo ga podijelili na train i test set. Prije procesiranja podataka, zvukove smo isjekli na duljinu od jedne sekunde, downsamplati na 22.05kHz, downmixali na mono te normalizirali sa RMS. Nakon toga smo koristili različite transformacije poput spektrograma, kromagrama i mfcc-a kako bi izvukli značajke iz zvukova te ih koristili kao inpute naše mreže.

Arhitektura mreže sastoji se od nekoliko konvolucijskih slojeva koji izvlače značajke iz zvukova, te od nekoliko gustih slojeva koji donose klasifikacijsku odluku. U mreži smo koristili različite aktivacijske funkcije, kao što su LeakyReLU i sigmoid, te MaxPooling i Dropout slojeve kako bi spriječili overfitting. Na kraju smo dobili mrežu s ukupno 9,628,811 parametara, od kojih je bilo 9,624,587 trenirajućih parametara.

Testirali smo našu mrežu na IRMAS test setu te smo dobili F1 score od 0.6, što je u skladu sa drugim state-of-the-art radovima u klasifikaciji instrumenata.

U zaključku, ovaj rad predstavlja pristup za klasifikaciju zvukova instrumenata koristeći duboku konvolucijsku neuronsku mrežu. Naš pristup pokazao se uspješnim na IRMAS datasetu, te bi se mogao primijeniti i na druge slične probleme klasifikacije zvukova. Potencijalna unaprijeđenja našeg pristupa uključuju korištenje većih skupova podataka za trening mreže, optimiziranje hiperparametara te korištenje novijih arhitektura neuronskih mreža.