УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Урош Стегић

АУТОМАТСКА ЖАНРОВСКА КЛАСИФИКАЦИЈА ПЕСАМА ТЕХНИКАМА ДУБОКОГ УЧЕЊА

мастер рад

Београд, 2023.

Ментор:

др Младен Николић, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Александар Картељ, доцент
Универзитет у Београду, Математички факултет

др Јована Ковачевић, доцент
Универзитет у Београду, Математички факултет

Датум одбране: _____

Садржај

Глава 1

Увод

Оринъачке фруле, дувачки музички инструменти произведени од костију животињског порекла, датирају из периода горњег палеолита и представљају најстарије откривене музичке инструменте [5]. Примерци ове фруле су пронађени у археолошким налазиштима у подручју Швапске Јуре у југоисточној Немачкој. Поновно датирање предмета пронађених на овим налазиштима методом радиоактивног угљеника је показало да су предмети пронађени конкретно у налазишту Гајзенклустал (нем. *Geissenklösterle*) старији од 40000 календарских година што фрулу пронађену на овом налазишту чини де факто најстаријим познатим музичким инструментом [11]. Фрула приказана на слици 1.1 поседује три отвора што јој омогућава свирање четири различита тона док се додатни призвуци могу добити различитим техникама дувања [5].

Слика 1.1: Оринъачка фрула израђена дубљењем радијалне кости лабуда

Од горњег палеолита до модерног доба, музика је имала разне улоге: ритуалне, уметничке, забавне итд. Током времена су се паралелно развијали нови музички инструменти и нови музички стилови. Данашњи инструменти су до-

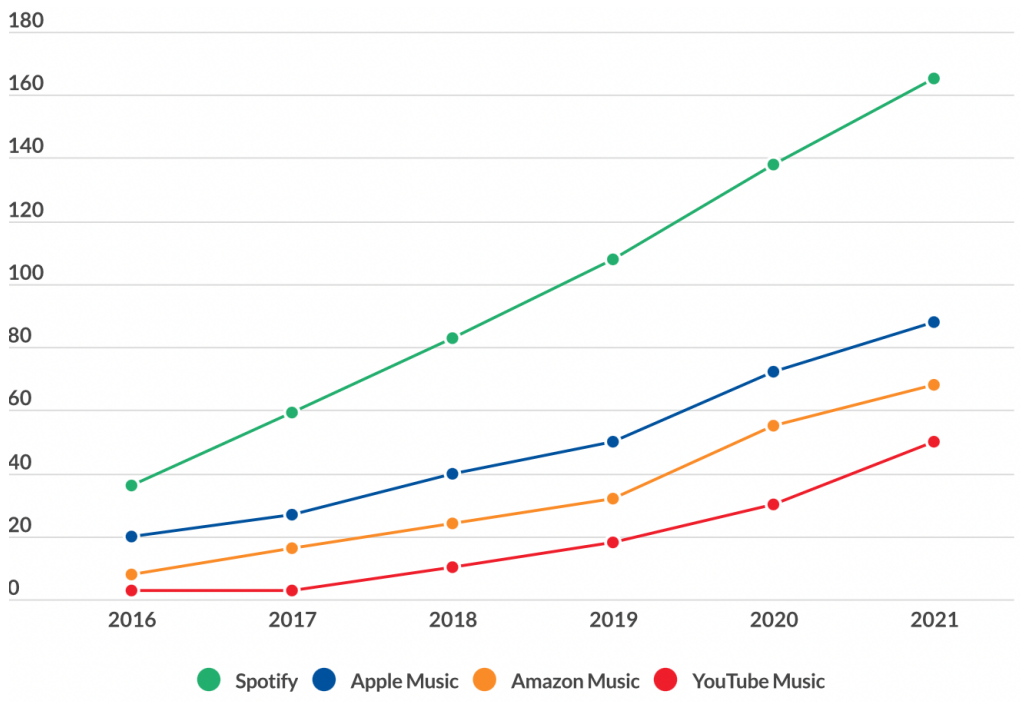
ста комплекснији од Оријачке фруле, а број музичких праваца непрестано расте.

Жанрови у музици немају формално дефинисан начин по коме се разликују једни од других већ се разликовање жанрова имплицитно описује кроз стилске сличности међу ауторима. На пример, репетитивна музика уз коју вокал има форму убрзаног наратива је стилска одредница која карактерише жанр „Хип Хоп". Наравно, овакав опис тог жанра јесте банализација тог музичког правца, јер је развој „Хип Хоп" музике пре свега културолошки феномен мотивисан положајем Афроамериканаца унутар Сједињених Америчких Држава. Овим се само илуструје комплексност коју разликовање музичких жанрова носи са собом. Прихваћену дефиницију музичких жанрова је понудио италијански музичар и музиколог Франко Фабри: „Музички жанр је скуп музичких догађаја чији правац је управљан коначним скупом социјално прихваћених правила" [8]. Поред тога што је дефиниција жанрова овако неформално формулисана, двадесет први век је донео нову комплексност - композитне жанрове (енгл. *fusion*) које карактерише субјективно самоопредељење аутора.

Средином деветнаестог века, конструкцијом фоноаутографа, а потом и фонографа (грамофона), звучни запис постаје могуће сачувати на физичком носачу звука – на грамофонској плочи. Запис звука и развој радио технологија су омогућиле појављивање првих радио станица почетком двадесетог века. Музика овим постаје један од распрострањенијих видова уметности и забаве. Велика распрострањеност музичког садржаја неминовно доводи и до његове монетизације и великих пословних прилика. Први пут се појављују комерцијални музички састави који производе широко прихваћену музику. Дугометражне композиције се замењују кратким нумерама које по форми више подсећају на фолклорне песме. Појављују се концепти попут музичких албума и хитова (једно-сезонских кратких песама које типично добију велику популарност у кратком временском року) и постепено се око музичке индустрије формира индустрија забаве.

Појавом интернета, тј. његовом техничком зрелошћу се појављују платформе за емитовање музике путем интернета (енгл. *streaming platforms*) као што су Епл Мјузик (енгл. *Apple Music*), Спотифај (енгл. *Spotify*), Дизер (енгл. *Deezer*) и друге. Битна разлика оваквог типа пласирања садржаја у односу на емитовање од стране радио станица се огледа у томе што не постоји ви-

ше програмска шема коју оформљава уредник програма, већ корисници сами бирају садржај који желе да конзумирају. Други аспект тога је чињеница да сваки корисник сада конзумира различит садржај насупрот радио емитовању где су сви слушаоци једног радија добијали исти садржај. Ове разлике су резултовале тиме да су овакви сервиси кренули развијати сопствене системе препорука како би своје кориснике додатно мотивисали да наставе коришћење сервиса. Системи препорука се могу заснивати на самим сличностима међу корисницима у смислу преференција ка сличним ауторима. Модерни системи своје препоруке заснивају на анализи самог звучног сигнала и проналажењу корисних атрибута тог сигнала који се могу искористити за проналажење сличних песама. Област рачунарства која се бави овим проблемима се зове истраживање музичких података (енгл. *music information retrieval, MIR*), у даљем тексту ИМП.



Слика 1.2: Број корисника платформи за емитовање музике изражен у милионима [6]

Британска компанија Соко Медиа (енгл. *Soko Media*) је 2022. године спровела истраживање пословања компанија које нуде услуге емитовања мултимедијалног садржаја [6]. У периоду од пет година су све компаније обухваћене

овим истраживањем забележиле велики раст многих параметара. На слици 1.2 је приказан пораст броја претплатника у периоду пет година. Платформа Спотифај, која се најбоље котира по броју претплаћених корисника, пријављује обим музичког садржаја од 80 милиона песама [1]. Оволика размера података представља велики изазов у њиховој обради.

Један од проблема који припадају ИМП је класификација музике на жанрове. Како је класификација песама (макар у некој основној форми која не захтева софистицирано доменско знање) релативно лак посао за човека, а сама формална дефиниција проблема не постоји, то се начин решавања проблема више заснива на статистичким методама. Последњих неколико деценија су технике машинског учења постале кључни део вештачке интелигенције и по својој природи веома добро одговарају оваквом типу проблема. У овом раду је истражено неколико популарних приступа класификацији музике на жанрове при чему су решења прилагођавана потпуно новом и веома обимном скупу података – Слободна музичка архива (енгл. *Free Music Archive*) [7].

У овом раду ће бити имплементирана три водећа приступа класификацији музике. Сви приступи су имплементирани „од нуле" и засновани су на коришћењу Слободне музичке архиве. Перформансе које су ови приступи постигли су мерени користећи контролни подскуп овог скупа који није учествовао у развоју самих решења. Резултати овог рада су нешто слабији у односу на резултате које су пријавили аутори ових решења. Разлика у перформанси је очекивана обзиром да су подаци који су коришћени у изградњи решења потпуно другачији, а додатно, недостатак рачунарских ресурса је велики ограничавајући фактор у изградњи оваквих решења.

Глава 2

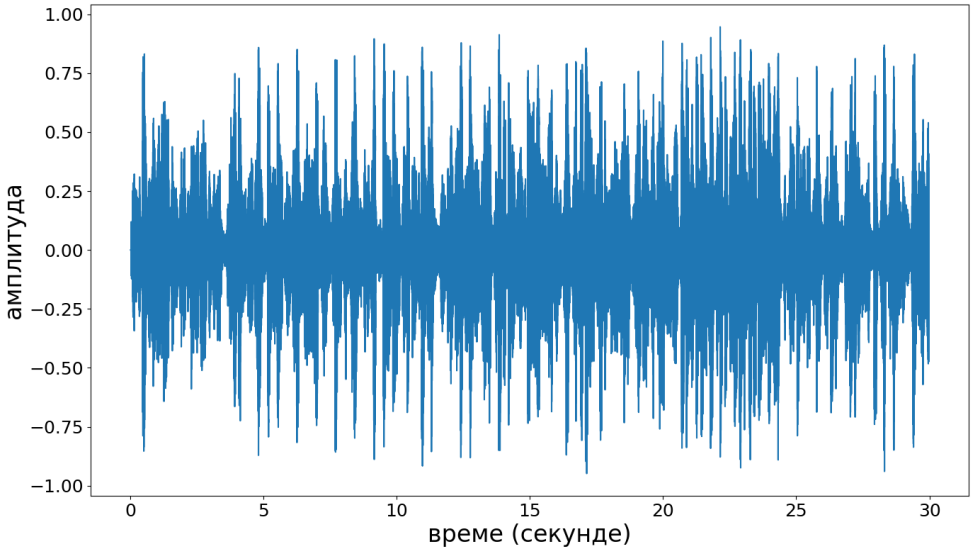
Анализа и обрада звука

Ово поглавље представља увод у модерну обраду и анализу звука и као такво се бави природом звука. Требало би да предочи одговоре на питања: „Чему служи анализа звука?", „На које начине се може обрадити звук?", „Које корисне информације звук носи?" и сл. Тема рада се односи на анализу музике у аудио сигналу, али је корисно осврнути се и на остале примене обзиром на то да су методе анализе и обраде веома сличне.

2.1 Физичке карактеристике звука и начин његовог чувања

Сам звук представља механички талас који настаје вибрацијом извора звука. Ове вибрације периодично мењају притисак флуида унутар којег се звук шири (нпр. ваздуха) док људски слушни апарат детектује ове осцилације притиска и преноси их на слушни нерв што даље мозак интерпретира као звук. На слици 2.1 је приказан звучни талас, при чему хоризонтална оса одговара времену, док вертикална оса представља амплитуду таласа.

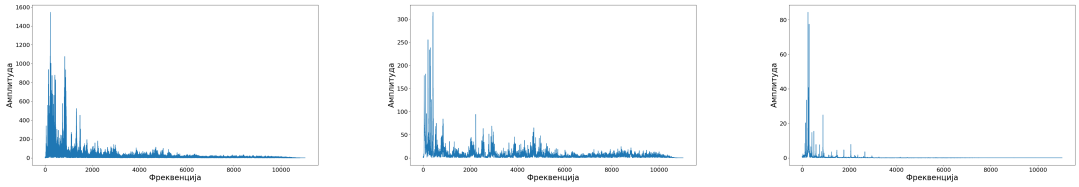
Један од првих носача звука, грамофонска плоча, садржи спирално урезане жљебове по којима се креће грамофонска игла. Неравнине на жљебовима представљају запис звука, а игла крећући се по жљебовима вибрира услед тих неравнина. Ове вибрације се преносе на магнет поред којег су постављена два калема. Вибрација магнета индукује електричну енергију унутар калемова чији поларитет и интензитет је директно управљан вибрацијом магнета. Сигнал из калема се преноси до уређаја који појачава овај сигнал и преноси га



Слика 2.1: Визуелна репрезентација звука у временском домену

даље до звучника. Унутар звучника, примљени сигнал се преноси до бакарног калема који индукује привремено магнетно поље које чији су поларитет и интензитет посредно условљени вибрацијама првог магнета. Унутар калема се налази магнет који се, реагујући на промене у магнетном пољу, креће, односно вибрира унутар жљеба. Овај централни магнет је залепљен за мембрану звучника (мембрана се често назива и дијафрагма) те своје механичко осциловање преноси на њу чије осциловање ствара механичке импулсе у ваздуху што људско ухо перципира као звук. Дакле грамофон механичко осциловање игле претвара у осциловање електричне струје, а затим ово осциловање даље претвара у механичко осциловање мембране звучника што генерише звук.

Како запис аналогног сигнала у рачунару није могућ, овакав сигнал се дигитализује тако што се он узоркује равномерно у распоређеним интервалима. Најквист-Шенонова теорема гарантује могућност реконструкције сигнала уколико је фреквенција узорковања макар дупло већа од највеће фреквенције у сигналу. Највећа фреквенција није позната током снимања, односно узорковања, али је познато да човек не може детектовати фреквенције веће од 20kHz, што би значило да је довољно узорковати било којом фреквенцијом већом од 40kHz. Технички преговори међу водећим компанијама у свету аудио технике су довели до стандардизације фреквенције узорковања на фреквенцију од 44.1kHz.



(а) Цела песма (б) Прва секунда песме (в) Последња секунда

Слика 2.2: Фуријеова трансформација различитих делова

2.2 Мотивација и примене

Анализа звука подразумева широк спектар примена. Поред очигледних примера попут креирања музичког садржаја, обрада звука се примењује у филмској индустрији, мобилној телефонији, технологијама попут активног поништавања буке (енгл. *active noise canceling*) и другима. Са друге стране, анализа звука је ширу индустријску примену нашла у производима новијег датума. Препознавање изговорених команди, препознавање изговореног текста, детекција активационих речи (енгл. *wake-up word detection*) су стандардни проблеми решавани анализом говора.

Алгоритам претраге музичких библиотека је популаризован апликацијом Шазам (енгл. *Shazam*). Основна идеја алгоритма је дефинисање отиска песме (енгл. *song fingerprint*) који се добија тако што се Фуријеова трансформација звучног сигнала подели на предефинисане сегменте и затим се за сваки сегмент изабере фреквенција највеће амплитуде. Базу података која чува овакве отиске је могуће веома брзо претражити што је огромна вредност коју овај алгоритам има [30]. Наравно, како снимак упита којим се претражује база није снимљен у студијским условима, оваква претрага не може увек пронаћи тачну песму. Претрага базе се може заменити филтрирањем отисака у бази и онда накнадним, мало комплекснијим, механизмима анализе звука доћи до правог резултата.

Технике анализе звука такође налази примену у тектоници где се звучним и ултразвучним таласима врши анализа земљишта, његова густина, детекција подземних вода, тражење руда итд. Сеизмологија такође проналази примену ових техника за рану детекцију земљотреса што је можда једна од племенитијих примена.

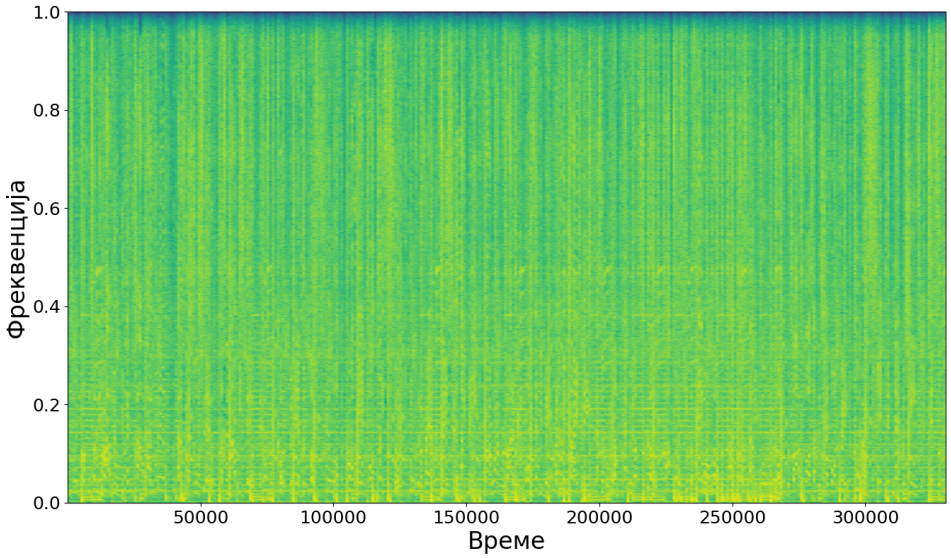
2.3 Спектрограми

Сваки сигнал се може деконструисати на линеарну комбинацију синуса и косинуса одговарајућих фреквенција. Коефицијенти линеарне комбинације се називају Фуријеови коефицијенти, а оваква трансформација Фуријеова трансформација. Применом Фуријеове трансформације улазни сигнал се из временског домена пресликава у фреквенцијски домен што омогућава лакшу анализу и обраду звука. На пример, шум из сигнала се може ублажити уклањањем свих фреквенција чије су амплитуде веће од задатог прага. За овако нешто је потребно извршити Фуријеову трансформацију сигнала, филтрирати резултујуће фреквенције и на крају инверзном Фуријеовом трансформацијом добија се поново аудио сигнал са умањеним шумом. Наравно, модерни приступи уклањања шума су далеко комплекснији и дају боље резултате.

Локална својства аудио сигнала често нису идентична у различитим деловима тог сигнала. На примеру музике се може видети како се фреквенције мењају од строфе до строфе, али чак и од такта до такта. Овакве варијације у сигналу отежавају његову анализу. На слици 2.2 се може уочити оваква природа аудио сигнала. Фуријеови коефицијенти израчунати над целим сигналом не носе довољно информација неопходних за било какво резоновање. Уместо тога, често се Фуријеови коефицијенти рачунају над одсечцима сигнала једнаких дужина. Сигнал се подели у n једнаких преклапајућих или дисјунктних целина, затим се над сваком целином израчунавају Фуријеови коефицијенти. Добијене вредности се ређају у матрицу тако што ће i-та колона матрице представљати амплитуде Фуријеових коефицијената i-те целине. На слици 2.3 је приказана визуелизација спектрограма топлотном мапом (енгл. *heatmap*).

2.4 Мелодијска скала сигнала

Перцепција стимуланса коју човек детектује својим чулима је пропорционална логаритму интензитета самих стимуланса [9]. Због овога се каже да су људска чула по природи логаритамска. На исти начин се перцепција висине тона зависи од фреквенције тог тона. Из тог разлога се приликом анализе музике ређе користе фреквенције добијене Фуријеовом трансформацијом већ се ове фреквенције пресликају на мелодијску скалу (енгл. *melscale*) следећом трансформацијом [19]:



Слика 2.3: Спектрограм

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

Овако трансформисан сигнал такође остаје у фреквенцијском домену, али се ове фреквенције, по конвенцији, не мере у херцима већ у меловима. Трансформација спектрограма мелодијском скалом се назива мелодијски спектрограм (енгл. *mel-spectrogram*) и често се користи у анализи музике и обради говора.

Честа операција која се користи у анализи звука је логаритмовање Фуријеове трансформације сигнала јер, попут висине тона, људска перцепција гласноће је такође логаритамска. Поновном Фуријеовом трансформацијом овако обрађеног сигнала се добија сепструм (енгл. *cepstrum*), док се сепструм добијен над мелодијском скалом назива сепструм мелодијских фреквенција (енгл. *mel-frequency cepstrum*). Корисна ствар код сепструма је његова способност да јасно раздвоји апстрактне карактеристике звука попут боје и висине тона.

2.5 Традиционални приступи класификацији звука

Традиционална класификација аудио сигнала се заснива на изградњи апстракција сигнала у виду проналажења његових атрибута, затим дефинисању скупа правила која на основу тако изграђених атрибута додељују сигналу једну од дефинисаних класа. Типично овакви приступи лоше генерализују, осетљиви су на промену окружења, позадински шум, промену опреме којом се врше мерења и сл.

Илустративни пример би могао бити класификација птица на основу звука којег птице испуштају. Преслушавањем података можемо приметити да у звуку који испушта велика сеница доминирају високе фреквенције, често испушта исти звук неколико пута узастопно без велике паузе између, док са друге стране звук који испушта обичан звиждак је мало дубљи, такође је сличног ритма међутим поседује веће паузе између узастопних гласова. На основу овог посматрања можемо дефинисати једноставан класификатор који има два атрибута: први атрибут би био просечна фреквенција у деловима сигнала који поседују неки звук, док би други атрибут био мало сложенији. Прво бисмо груписали делове сигнала који поседују звук, затим бисмо пребројали колико таквих делова има у некој јединици времена (рецимо током пет секунди). Ова два атрибута би могла да нам одреде о којој птици је реч.

Према домену над којим се граде, атрибуте можемо поделити на спектралне и временске. Временски атрибути се добијају обрадом изворног сигнала, док се спектрални добијају анализом фреквенција које се налазе у сигналу тј. спектрални атрибути настају обрадом Фуријеове трансформације сигнала.

Неки од често коришћених временских атрибута [20]:

- Стопа преласка нуле (енгл. *zero-crossing rate*) представља стопу којом сигнал мења знак.
- Средњеквадратна амплитуда (енгл. *RMSE amplitude*) је средњеквадратна вредност амплитуде, представља просечну снагу сигнала и користи се да опише општу гласноћу звука.
- Временски центроиди (енгл. *temporal centroids*) представљају просеке амплитуда исечака аудио сигнала отежане временском компонентом и

тиме описују како је енергија сигнала расподељена кроз време. Тежински фактор свакој амплитуди је представљен тренутком у ком се та амплитуда десила (нпр. број секунди релативан у односу на дати исечак). Овај атрибут је нарочито користан у идентификацији инструмената пошто једна од основних физичких карактеристика боје звука јесте динамика расподеле енергије кроз време.

- Аутокорелација (енгл. *autocorrelation*) представља корелацију сигнала са временски помереном верзијом истог тог сигнала. Уместо да се аутокорелација дефинише као функција помераја, у пракси се често конструише само неколико конкретних атрибута дефинисаних различитим померајем. Аутокорелацијом се детектују периодичности у сигналу.

Примери атрибута добијених над фреквенцијским доменом сигнала [20]:

- Спектрални центроиди (енгл. *spectral controids*) представљају просечне фреквенције у (унапред дефинисаним) одсечцима сигнала, а поред временских центроида се такође користе при карактеризацији боје звука.

- Опсег фреквенција (енгл. *spectral bandwith*) се рачуна као разлика између највеће и најмање фреквенције које су присутне у сигналу и често се користи за категоризацију фамилије инструмената по висини тонова које праве (нпр. бас, баритон, сопрано итд.).

- Фреквенцијски флукс (енгл. *spectral flux*) је мера промене доминантних фреквенција кроз време

- Коефицијенти мелодијског сепструма (енгл. *mel-frequency cepstral coefficients, MFCC*) су веома корисни атрибути јер коефицијенти ниских фреквенција одговарају боји звука у музици (односно фонемама у обради говора) док коефицијенти виших фреквенција одговарају висини тона.

Глава 3

Машинско учење

Машинско учење је област рачунарства која се бави изградњом система заснованих на индуктивном закључивању. У великом броју случајева није тешко написати софтвер од којег се очекује да одговори на неки скуп формално дефинисаних захтева попут информационих система, банкарског софтвера, и слично. Ипак, традиционалан приступ развоју софтвера јако тешко или чак никако не може решити проблеме као што су препознавање лица, аутоматско управљање возилима, машинско превођење природних језика итд.

Типично се технике машинског учења заснивају на употреби сакупљених података на основу којих би систем који се развија могао сам научити карактеристике тих података, те на основу њих доносио одлуке које се од њега траже. Како овакав приступ доста подсећа на концепт учења код људи и животиња, одатле и назив области. Технике машинског учења су у многоме замениле раније приступе развоју вештачке интелигенције и данас се под термином вештачке интелигенције често крије машинско учење. Многи аутори ове термине користе наизменично.

Количина мултимедијалног садржаја се популаризацијом интернета и социјалних мрежа умногостручила, што заједно са порастом рачунске моћи представља погодно тло за развијање машинског учења. Употреба графичког процесора представља једну од прекретница ове области [24]. Друга прекретница у машинском учењу представља рад из 2012. године који је надмашио људске перформансе у задатку класификације слика где се тражило од система да сваку слику означи тачно једном од унапред дефинисаних класа.

Ово поглавље ће представити машинско учење, садржаће опис коришћених метода и послужиће као основа за разумевање теме рада.

3.1 Основни концепти машинског учења

Машинско учење као дисциплина се бави развојем система који нису експлицитно програмирани већ на основу података уче да решавају задатак. Типично модел представља оцену расподеле података, при чему су параметри те расподеле предмет „учења". Ови параметри се бирају тако да минимизују одговарајућу функцију грешке. Процес проналажења најбољих параметара се зове обучавање модела и врши се алгоритмима оптимизације, а подскуп података којима се модел обучава се назива скуп за обучавање.

Основни метод оптимизације који се користи у машинском учењу је алгоритам градијентног спуста. Он припада фамилији итеративних, градијентних метода. У основи, овај метод функционише тако што у првом кораку почне од произвољне тачке, израчуна градијент функције у тој тачки, затим нову тачку добије тако што се од постојеће тачке помери супротно од смера градијента и то скалираним интензитетом градијента. Коефицијент скалирања се често означава са α и назива се кораком учења. Овај корак се затим понавља докле год није испуњен претходно дефинисан критеријум заустављања.

Algorithm 1 Алгоритам градијентног спуста

function GradientDescent(f, α)
  $x_0 \sim \mathcal{U}[Domain(f)]$ ▷ Иницијализација се може разликовати
  $i \leftarrow 0$
  repeat
    $x_{i+1} \leftarrow x_i - \alpha \frac{\nabla f}{\nabla x}(x_i)$
    $i \leftarrow i + 1$
  until converged
end function

Линеарна регресија је метода која се користи за предвиђање нечега што је по природи реална променљива. На пример може се захтевати од модела да предвиђа цену некретнине на основу неких других нумеричких карактеристика попут броја соба, квадратуре и удаљености од центра. Модел линеарне регресије ће у овом случају изгледати овако: $f(x_1, x_2, x_3) = w_0 + w_1x_1 + w_2x_2 + w_3x_3$ при чему су w_i параметри модела, док су x_i атрибути који представљају поменуте карактеристике некретнина. Једноставности ради, модел се може представити векторски на следећи начин: $f(\mathbf{x}_i) = \mathbf{x}_i^T\mathbf{w}$, при чему вектор $\mathbf{x}_i$ одговара i-тој инстанци. Нека је неколико примера кућа на основу којих модел може да учи представљено паровима $(\mathbf{x}_i, y_i)$ где је $\mathbf{x}_i$ вектор сачињен од

вредности атрибута i-те куће док је y_i циљна променљива која представља цену i-те куће. Параметри $\mathbf{w}$ се одређују тако да минимизују следећу функцију грешке:

$$L(\mathbf{w}; \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$$

Типови учења

На основу механизама учења и врсте скупа података, типично се разликују:

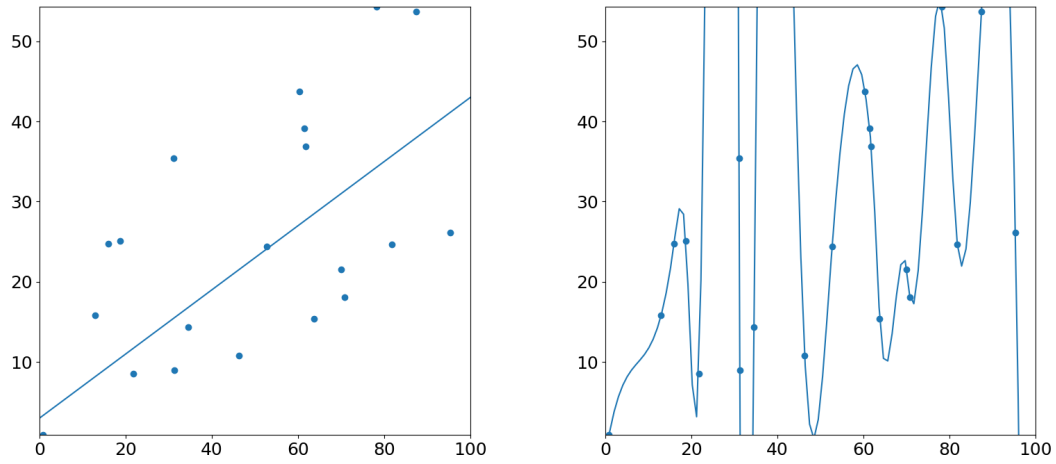
- надгледано учење
- ненадгледано учење
- учење поткрепљивањем

Надгледано учење представља врсту проблема где су подаци унапред означени (на било који начин, али типично људским трудом), дакле проблеми у којима је вредност циљне променљиве унапред позната у скупу за обучавање. Линеарна регресија је пример надгледаног учења јер је тачна вредност цене куће била доступна током обуке модела.

Ненадгледано учење се бави решавањем проблема у којима подаци нису унапред означени, те вредност циљне променљиве није позната током обуке модела, тј. циљна променљива као таква не постоји у проблемима ненадгледаног учења. Типично се овакви модели користе за откривање структуре података са којима се ради.

Учење поткрепљивањем је трећа велика област машинског учења. У стандардној поставци учења поткрепљивањем постоји агент који интерагује са некаквим окружењем. У сваком моменту се агент налази у одређеном стању. Предузимањем акције, агент прелази из једног стања у друго и овим преласком од окружења добија награду или казну. Циљ агента је да научи коју акцију да предузме у ком стању на начин тако да максимизује укупну награду. Класични примери употребе учења поткрепљивањем су играње видео игара, роботика, аутоматско управљање, и слично.

Поред поделе засноване на врсти проблема односно типу скупа података на основу којих се обучава модел, битна подела самих модела је на дискриминативне и генеративне. Дискриминативни модели моделују зависност циљне



(а) Подприлагођавање (б) Преприлагођавање

Слика 3.1: Прилагођавање модела

променљиве од атрибута кроз одређивање условне расподеле $p(y|\mathbf{x})$ док генеративни модели покушавају да моделују цео скуп података кроз одређивање заједничке расподеле $p(\mathbf{x}, y)$ односно скраћено – само $p(\mathbf{x})$. Једноставном применом Бајесове теореме се генеративни модел може свести на дискриминативни. Типично, обзиром да познају целу расподелу у подацима, генеративни модели су у могућности да генеришу нове, до сада невиђене, инстанце па одатле и добијају свој назив.

Преприлагођавање

Током обуке модела, лако се може догодити да модел веома добро опише неке специфичности скупа којим је модел обучаван и тиме не успеју да генерализују. Овакви модели дају јако лоша предвиђања, па је корисно избећи преприлагођавање модела. Проблем потприлагођавања је супротан проблем, другим речима, то је ситуација у којој модел у фази обуке није успео да се прилагоди подацима довољно. На слици 3.1 се могу видети примери потрприлагођавања (лево) и преприлагођавања (десно).

Класификација

Један од значајнијих проблема који се моделују дискриминативним моделима је свакако проблем класификације. Поставка класификације је таква да је унапред дефинисан скуп од две или више класа којима свака инстанца може да припадне, при чему свакој инстанци може бити додељена тачно једна класа. Типични примери класификације су одређивање малигности тумора на основу његових физичких карактеристика или одређивање расе пса на основу слике.

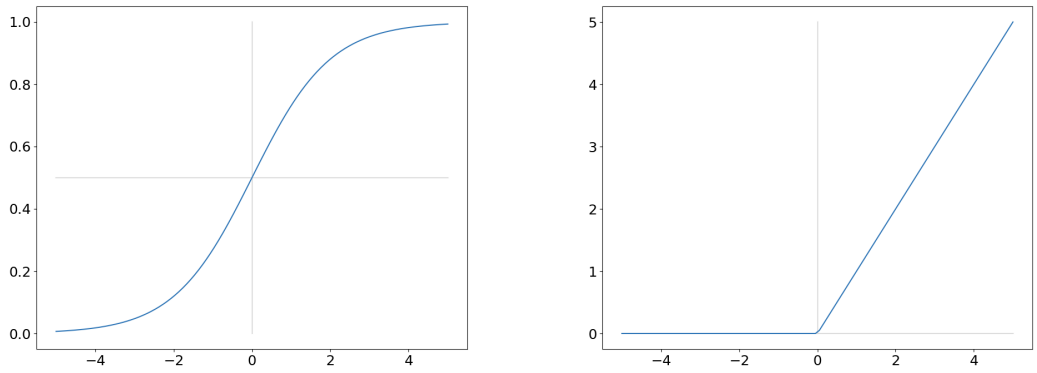
Као што линеарна регресија моделује вредност произвољне реалне променљиве, тако логистичка регресија моделује вероватноћу за жељеним исходом. Прецизније, код модела линеарне регресије, вероватноћа $p(y|\mathbf{x})$ је по претпоставци нормално расподељена, док се код логистичке регресије претпоставља да ова вероватноћа има Бернулијеву расподелу. Основни модел логистичке регресије и функција грешке су представљени следећим формулама:

$$p(y|\mathbf{x};\mathbf{w}) = \sigma(\mathbf{x}^T\mathbf{w})^y(1-\sigma(\mathbf{x}^T\mathbf{w}))^{1-y}, \qquad \sigma(z) = \frac{1}{1+e^{-z}} \tag{3.1}$$

$$L(w) = -\sum_{i=1}^{n} y_i\sigma(\mathbf{x}^T\mathbf{w}) + (1-y_i)(1-\sigma(\mathbf{x}^T\mathbf{w})) \tag{3.2}$$

Функција $\sigma(z)$ у изразима 3.1 и 3.2 представља сигмоидну функцију, а њен график се може видети на слици 3.2а. Ова нелинеарна функција је ограничена на интервалу $(0, 1)$ и монотоно је растућа на целом домену. Функција у изразу 3.2 се назива унакрсна ентропија (енгл. *cross entropy*) и стандардно се користи као функција грешке у проблему класификације.

Поред класификације постоји и задатак означавања (енгл. *tagging*). Циљ означавања је придруживање једне или више ознака (из неког унапред дефинисаног скупа ознака) свакој инстанци. Главна разлика у односу на класификацију је та што код класификације свакој инстанци додељујемо тачно једну класу (ни мање ни више) док код означавања можемо свакој инстанци доделити различит број ознака. Један начин да се инстанце означе једном од n ознака је коришћењем n независних класификатора.



(а) Сигмоидна функција (б) Исправљена линеарна јединица

Слика 3.2: Активационе функције

3.2 Неуронске мреже

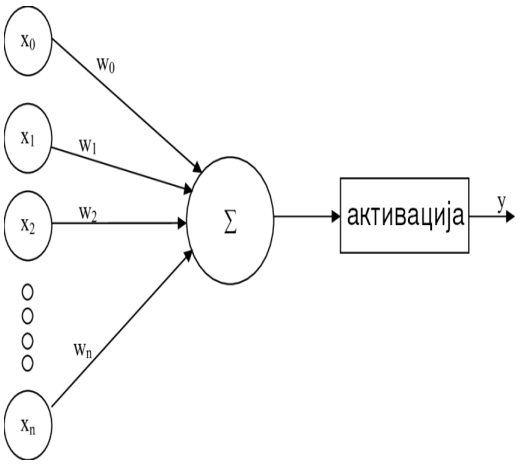
Неуронске мреже представљају класу модела машинског учења који се граде увезивањем тзв. неурона, где сваки неурон процесира и прослеђује своје улазне информације следећем неурону. Типично, сваки неурон је један линеаран модел чији резултат се пропушта кроз нелинеарну функцију. Модел логистичке регресије је добар пример једног неурона.

Потпуно повезане неуронске мреже

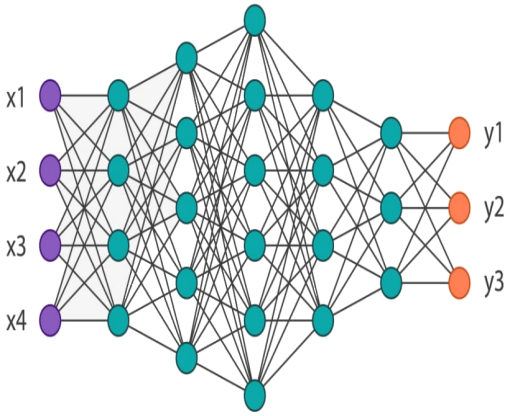
Основни облик неуронских мрежа представља потпуно повезана неуронска мрежа. Код овог типа мрежа, неурони се групишу у слојеве при чему једна неуронска мрежа може имати један или више слојева не рачунајући слој који представља улазне податке нити слој који представља резултат неуронске мреже. Ови унутрашњи слојеви мреже се такође називају скривеним слојевима. Сваки неурон се може посматрати као линеарни модел који поседује свој вектор тежина[1]. Излаз неурона се добија линеарном комбинацијом параметара тог неурона са излазима свих неурона претходног слоја. На слици 3.3 је приказан појединачан неурон (лево) и једна потпуно повезана неуронска мрежа (десно).

Овако дефинисан модел неуронске мреже је линеаран пошто се крајњи

[1]До сада је био коришћен термин *параметар*, ова два термина су синоними. У контексту неуронских мрежа се у литератури чешће среће израз тежина.



(а) Неурон



(б) Потпуно повезана неуронска мрежа са пет скривених слојева

Слика 3.3: Схематски приказ потпуно повезане неуронске мреже

излаз добија као композиција линеарних функција која је сама по себи такође линеарна. Како линеарни модели нису у стању да опишу нелинеарне правилности у подацима, оваква дефиниција неуронске мреже се модификује нелинеарном трансформацијом излаза сваког неурона. Функције примењене над излазима неурона се називају активационим функцијама.

Активације

Једна од типичних активационих функција је сигмоидна функција дефинисана у претходном делу једначином 3.1. Када се користи у излазном слоју који садржи један неурон, неуронска мрежа тада моделује условну вероватноћу расподељену Бернулијевом расподелом, само за разлику од логистичке регресије може моделовати нелинеарну зависност ове вероватноће од улазних података.

У случају када је потребно разликовати више класа, подаци се моделују категоричком расподелом. Тада се у излазном слоју користи функција меког максимума (енгл. *softmax*) дефинисана у изразу 3.3.

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}, \qquad i = 1, ..., C \tag{3.3}$$

Код унутрашњих слојева мреже, функција активације која се често користи је исправљена линеарна јединица (енгл. *rectified linear unit, ReLU*). Израз

3.4 даје дефиницију ове функције док слика 3.2б приказује график ове функције. Исправљена линеарна јединица није диференцијабилна у нули, међутим ово не представља проблем оптимизационим методама јер су ситуације да вредност неурона пред активацију буде баш нула веома ретке, а у случају кад се догоде може се додефинисати вредност градијента у нули што неће покварити процес обуке модела.

$$ReLU(z) = max(0, z) \tag{3.4}$$

Алгоритам пропагације уназад

Обучавање неуронске мреже функционише као обучавање до сада поменутих модела. Користећи технике нумеричке оптимизације попут градијентног спуста, параметри мреже се ажурирају користећи њихов градијент. Алгоритам за израчунавање овог градијента се назива алгоритам пропагације уназад. Како неуронска мрежа није ништа више до композиције функција, тако се и градијенти по параметрима рачунају правилом рачунања извода сложене функције (енгл. *chain rule*).

Прво је неопходно израчунати предвиђање, тј. добити резултате неуронске мреже. На основу добијених резултата и тачне вредности циљне променљиве се може израчунати функција грешке. Пропагација уназад почиње рачунањем градијента функције грешке којим се ажурирају тежине последњег слоја. Овај градијент се даље пропагира на претходни слој где фигурира у израчунавању градијента тог слоја правилом извода сложене функције. Овај процес се даље понавља све до улазног слоја.

3.3 Конволутивне мреже

Потпуно повезане неуронске мреже су типично велики модели гледано по броју параметара. Ипак, приликом процесирања визуелних и аудио информација се могу користити специфичности тог типа сигнала. На пример, лопта исто изгледа на ком год делу слике да се налази па је с тога непотребно имати посебне неуроне који гледају само одређене делове слике.

Конволутивне мреже се такође граде у хијерархијском маниру градећи слојеве мреже који се надовезују један на други. Обучавање конволутивних

мрежа функционише идентично као обучавање потпуно повезаних неуронских мрежа – градијент се рачуна алгоритмом пропагације уназад, а затим се параметри модела ажурирају неком од варијација градијентног спуста. Основна два типа слојева који се срећу код конволутивних мрежа су конволутивни слој и слој агрегације.

Слој конволуције

Конволуцијом се постиже екстраховање информација из улазног сигнала те се на овај начин добија аутоматско грађење атрибута насупрот ручном одабиру атрибута описаних у одељку 2.5. У зависности од врсте сигнала разликују се једнодимензионе и дводимензионе конволуције. Фундаментално су обе врсте идентичне: улазни сигнал се подели у делове идентичних димензија, затим се сваки део сигнала конволуира истим филтером. Ови делови сигнала су ретко дисјунктни, а подела се контролише метапараметром који се често назива корак конволуције при чему је типична вредност корака – један. На тај начин се постиже ефекат као да филтер конволуције „клизи" кроз улазни сигнал. Израз 3.5 приказује операцију конволуције дела дводимензионог сигнала s филтером q при чему су димензије дела сигнала и филтера $m \times n$.

$$s * q = \sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij} q_{ij} \tag{3.5}$$

Након што се филтером q примени конволуција над свим деловима сигнала, резултати ових конволуција се спајају у један излазни сигнал који сада представља изграђен један атрибут. Применом различитих филтера се добијају различити атрибути. Један конволутивни слој мреже се састоји од произвољног броја филтера, а вредности у овим филтерима представљају параметре који се уче процесом обучавања модела. Излаз једног конволутивног слоја се добија тако што се улазни сигнал, односно улазни атрибути конволуирају свим филтерима тог слоја, затим се над резултатима примени функција активације и тако добијени атрибути се сложе у тензор. На пример, ако слој садржи p филтера и сваки филтер након конволуције и активације произведе матрицу димензије $m \times n$, крајњи излаз овог слоја ће бити тензор димензије $m \times n \times p$.

Слој агрегације

За разлику од конволутивних слојева, слојеви агрегације немају параметре већ искључиво служе како би компресовали информације генерисане конволутивним слојем. Овим видом упрошћавања атрибута се постиже боља генерализација, а поред тога се смањује рачунска комплексност модела што је веома важно код дубоког учења где су модели типично веома велики и где се обрађују велике количине података.

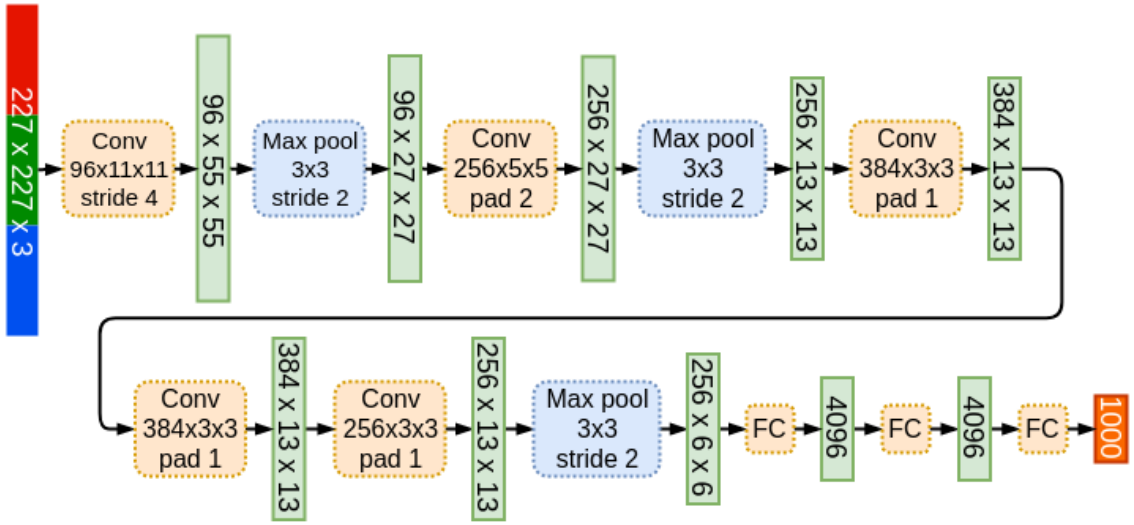
Начин примене слоја агрегације је сличан конволутивним слојевима. Улазни сигнал (односно атрибути добијени конволутивним слојем) се поново подели на једнаке делове (овога пута често дисјунктне делове), затим се над сваким делом примени скаларна функција агрегације. Избор ове функције је произвољан, а најчешћи избори су функција просека и функција максимума. Зато се најчешћи слојеви агрегације називају слој агрегације просеком (енгл. *average pooling layer*) и слој агрегације максимумом (енгл. *max pooling layer*).

Архитектуре

Уобичајен начин изградње конволутивних неуронских мрежа је наизменично слагање конволутивних слојева и слојева агрегације. Избор дубине мреже, димензија филтера, активационих функција, типова агрегационих слојева су метапараметри. За потребе класификације се често излаз последњег слоја пресложи у вектор који се затим користи као улаз потпуно повезане неуронске мреже која заправо обавља саму класификацију. Из тог разлога се у стандардној терминологији први део мреже који је сачињен од конволутивних и агрегационих слојева назива „екстрактор атрибута" (енгл. *feature extractor*) док се други део мреже сачињен од потпуно повезаних слојева назива класификатор (енгл. *classifier*). Слика 3.4 приказује неке од популарних архитектура конволутивних неуронских мрежа за решавање проблема класификација слика.

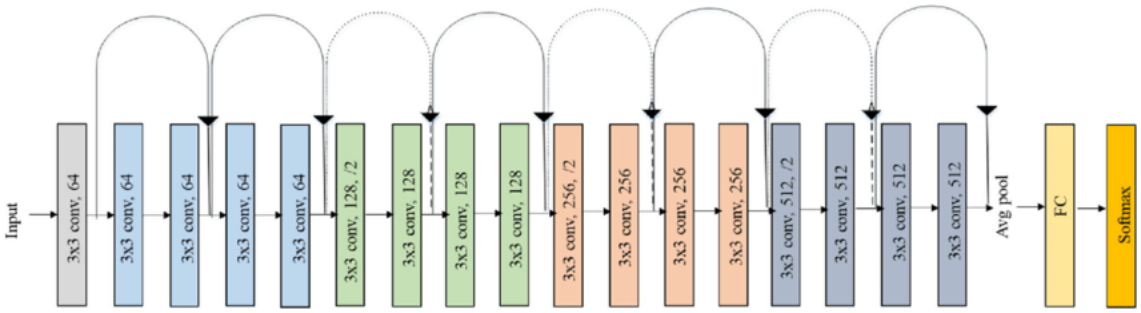
3.4 Евалуација модела

Како се машинско учење бави изградњом модела који се труде да генерализују правилности пронађене у подацима, тако је циљ евалуације модела да се процени како ће се модели понашати над новим подацима приликом



(а) AlexNet [13]

(б) VGG16 [25]



(в) Resnet18 [10]

Слика 3.4: Популарне архитектуре конволутивних мрежа

његове употребе. Приликом евалуације модела је веома важна педантност у методологији пошто наизглед ситне и суптилне грешке могу довести до лажних оцена.

Методологија избора и евалуације модела

Како су модели склони прилагођавању самим подацима уместо правилностима у тим подацима, вршити оцену квалитета користећи скуп коришћен за обучавање неће дати добре резултате. Уместо тога се често полазни скуп података дели на подскуп за обуку модела и подскуп за евалуацију модела.

Наредни проблем је избор метапараметара. Ако се за метапараметре бирају оне вредности које на тестном скупу дају најбоље резултате то значи да ти метапараметри нису оптимални метапараметри у општем смислу већ да су то најбољи метапараметри за тај тест скуп. Другим речима, мера квалитета добијена на овај начин није индикативна за понашање модела на новим подацима. Уместо тога се полазни скуп дели на три дела: подскуп за обучавање модела, подскуп за избор метапараметара (који се често назива валидациони скуп) и подскуп за мерење квалитета – тестни скуп.

Мере квалитета

Опис мера квалитета се односи на случај када се врши бинарна класификација, уопштење на више класа се типично постиже одговарајућим агрегацијама. Најпре се једна класа прогласи позитивном класом. Избор позитивне класе је произвољан и препуштен је интуицији. На пример, ако модел врши класификацију инстанци снимака тумора онда позитивна класа може бити да је инстанца тумора малигна. Најчешће коришћене мере квалитета модела који врше класификацију су прецизност (енгл. *precision*), одзив (енгл. *recall*) и површина испод РОК криве (енгл. $ROC\ AUC$). Прве две мере се заснивају на томе да се вероватноће које модел даје заокруже користећи неки праг. Формуле за рачунање прецизности и одзива су дате у изразу 3.6, при чему TP означава број инстанци позитиве класе које је модел класификовао позитивном класом (енгл. *true positive*), FP број инстанци негативне класе које је модел класификовао позитивном класом (енгл. *false positive*), FN број инстанци позитивне класе које је модел класификовао негативном класом (енгл. *false negative*) и коначно TN број инстанци негативне класе које је модел кла-

сификовао негативном класом (енгл. *true negative*). Опсег ове мере се креће у интервалу $[0, 1]$.

За разлику од прве две, површина испод РОК криве није осетљива на избор вредности прага. Ова мера представља вероватноћу да при насумичном избору двеју инстанци различитих класа, модел инстанци негативне класе придружи мању вредност. Вредност ове мере се рачуна се помоћу формуле у изразу 3.7, при чему C_1 и C_2 представљају скупове инстанци негативне и позитивне класе, тим редом, док $f(i_n)$ представља вредност коју модел додељује n-тој инстанци [18]. Опсег ове мере се креће у интервалу $[\frac{1}{2}, 1]$.

$$precision = \frac{TP}{TP + FP}, \qquad recall = \frac{TP}{TP + FN} \tag{3.6}$$

$$AUC = \frac{|\{(i_1, i_2) \mid i_1 \in C_1 \wedge i_2 \in C_2 \wedge f(i_1) < f(i_2)\}|}{|C_1| \cdot |C_2|} \tag{3.7}$$

Глава 4

Класификација музике дубоким учењем

Као и у другим областима, примат у обради и анализи звука су преузеле технике дубоког учења. Већина научних радова објављених 2022. године на водећим конференцијама попут ИСМИР (енгл. *International Society of Music Information Retreival, ISMIR*) и ИЕЕЕ (енгл. *Institute of Electrical and Electronics Engineers, IEEE*) одсека за акустику, говор и обраду сигнала су из области машинског учења.

Музички жанрови нису прецизно дефинисан појам, као што је описано и у секцији 1. Често припадност неком жанру буде и субјективна тј. сами уметници дефинишу своје жанрове и често и измишљају нове. Поред тога, неретко уметници стварају музику која има елементе неколицине жанрова. Због тога често системи за анализу музике не врше стандардну класификацију, већ обављају задатак означавања (енгл. *tagging*) где се свакој инстанци придружује скуп од неколико жанрова. Тако на пример нека песма може бити означена жанровима електро и рок ако поседује елементе оба ова жанра.

4.1 Скупови података

Велики изазов за прикупљање података из области анализе музике представљају ауторска права. Преузимање и умножавање музике није дозвољено, а куповина музике у обиму који је неопходан дубоким неуронским мрежама је веома скупа. У наставку следи опис најзначајнијих скупова података над којима су обучавани актуелни модели.

ГТЗАН

Симболични представник скупова података за жанровску класификацију музике је скуп ГТЗАН (енгл. *GTZAN*) [29]. Скуп је настао 2002. године, садржи хиљаду песама подељених у десет жанрова, при чему сваку песму физички репрезентује аудио запис у трајању од тридесет секунди. Сам скуп је превише малог обима како би се користио при обучавању модела, тако да је данас ГТЗАН је један од најкоришћенијих скупова података за евалуацију модела жанровске класификације музике [28].

Упркос његовој широкој примењености, овај скуп поседује недостатке: погрешне анотације, квалитет аудио записа и постојање дупликата [27]. Проблем са самим ознакама је значајан проблем – скоро десет посто песама је погрешно означено. Квалитет аудио записа не утиче превише на употребљивост скупа јер свега неколико песама има аудио запис који је оштећен до границе употребљивости. Дупликати са друге стране могу значајно утицати на квалитет евалуације. Препознате су три категорије дупликата: неке песме су комплетно дуплиране, неке песме потичу из истог оригиналног снимка, али немају потпуно преклапање и на крају постоје песме које су снимљене у неколико верзија (нпр. студијски снимак и снимак исте те песме изведен уживо).

Скуп од милион песама

Скуп од милион песама (енгл. *Milion Song Dataset, MSD*) [2] је настао 2011. године и једини је међу популарним скуповима који заправо не поседује ниједан аудио запис. Ова чињеница управо представља и један од његових највећих недостатака. Као што му име говори, овај скуп садржи податке о милион песама па по свом обиму представља један од највећих скупова података у овој области. Скуп је сачињен од унапред израчунатих аудио атрибута при чему су доступни и мета подаци о песмама.

Поред недоступности аудио записа, оригинални скуп је поседовао и проблем у подели података. Подела коју су аутори рада предложили је заснована на простом случајном узорку песама. На овај начин музика једног аутора може да се нађе и у подскупу за обучавање модела и у подскупу за евалуацију што доводи до пристрасне оцене квалитета. Нова подела са стратификацијом на основу аутора песама решава тај проблем [31].

Магна таг тјун

Први музички скуп података који је састављен за потребе означавања уместо уобичајене класификације је Магна таг тјун (енгл. *MagnaTagATune*) [14]. Скуп се састоји од 5405 песама трајања по 29 секунди. Музика долази од независних музичара (енгл. *indie music*), те квалитет продукције није нивоа професионалних продуцентских кућа. Саме ознаке на скупу су такође подложне шуму јер процес анотирања није имао ограничења (предефинисане класе). Интересантно је да је анотирање било сакривено у видео игри где су играчи прво означили песму, а затим би сваком играчу биле приказане ознаке другог играча како би тај играч проценио да ли су слушали исту песму. Ово је довело до шума у ознакама што овај скуп чини непрактичним за евалуацију модела.

МТГ Јамендо

МТГ Јамендо (енгл. *MTG-Jamendo*) је састављен 2019. године од стране истраживачке лабораторије МТГ из Барселоне [3]. Скуп садржи 55000 целих песама означених некима од 195 ознака. Ознаке у овом скупу нису искључиво жанровске већ су поред самог жанра означени инструменти и перципирано расположење које песма подстиче. Музика је компонована и изведена од стране независних аутора и објављена на платформи Јамендо за слободно преслушавање. Квалитет аудио записа је висок, али проблем који овај скуп података поседује је небалансираност жанрова.

Слободна музичка архива – CMA

Слободна музичка архива (енгл. *Free Music Archive, FMA*) је највећи скуп података који садржи аудио записе [7]. Настао је 2017. године од стране истраживачке групе при универзитету у Лозани. Сва музика овог скупа је снимљена за потребе скупа података и као таква је бесплатна за употребу у истраживачком раду. Цео скуп садржи преко сто хиљада песама, што збирно износи око 353 дана музике. Зарад лакше употребе, скуп је подељен у мање целине при чему је сваки скуп строги подскуп следећег: *мали*, *средњи*, *велики* и *потпуни*. Табела 4.1 приказује детаљније информације о овим подскуповима.

Музика снимљена за овај скуп је под лиценцом креативних заједничких добара (енгл. *Creative Commons*). Пошто је, као и код претходних скупова, снимљена од стране независних аутора остаје упитан квалитет као и његова репрезентативност у реалном свету. Скуп је означен са 163 жанра распоређених у хијерархијску таксономију, на пример из жанрова соул и Р&Б (енгл. *rythm & blues*) су изведени жанрови диско и фанк, док из фанка произилази жанр дубоки фанк (енгл. *deep funk*).

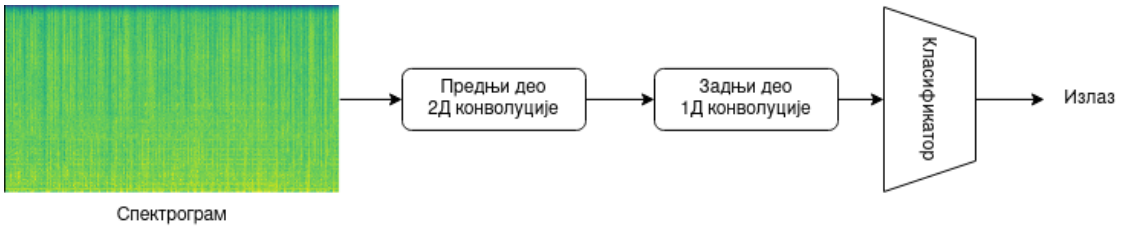
Подскуп	# песама	# жанрова	просечно трајање (s)	обим података (GB)
мали CMA	8000	8	30	7.4
средњи CMA	25000	16	30	23
велики CMA	106574	161	30	98
потпун CMA	106574	161	278	917

Табела 4.1: Подскупови CMA [7]

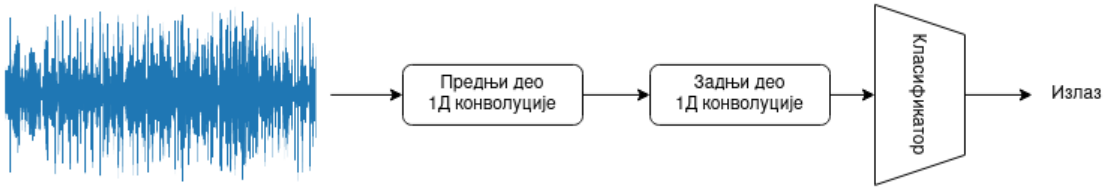
Модели који постижу најбоље резултате из области жанровске класификације музике нису обучавани нити евалуирани на CMA скупу. Чак, неретко нису евалуирани на истим скуповима што онемогућава њихово адекватно поређење. Један од доприноса овог рада представља обучавање постојећих модела над CMA скупом и њихова евалуација.

4.2 Модели

Типично, модели коришћени за класификацију музике по било ком својству, па тако и жанровима, су конволутивне неуронске мреже. Стандардна поставка конволутивне неуронске мреже за класификацију звука се састоји од три основна дела: предњи део мреже (енгл. *front-end*), задњи део мреже (енгл. *back-end*) и класификатор. Прва два дела су саграђена користећи конволутивне слојеве док се класификатор састоји од потпуно повезаних слојева. Разлог за раздвајање предњег и задњег дела мреже је у томе да би постојао део дељене архитектуре (задњи део и класификатор) у верзијама модела који раде директно над аудио сигналом и модела који раде над спектрограмима. Типично се задњи део састоји од једнодимензионих конволуција док се предњи део састоји од једнодимензионалних или дводимензионалних у зависности од тога да ли на улазу примају сиров аудио сигнал или спектрограме.



(а) Предњи део за спектрограме



(б) Предњи део за аудио сигнал

Слика 4.1: Стандардни елементи конволутивних мрежа за анализу звука

Поред техничке разлике, сматра се да постоји интерпретабилност слојева налик оној у конволутивним мрежама за класификацију слика мада је у случају музике то теже показати визуелно. Први слојеви мреже, односно предњи део мреже, током обуке модела науче да препознају локалне атрибуте акустике попут висине тона, темпа, ритма, боје звука итд. Задњи део мреже се типично фокусира на секвенце изграђених атрибута и тиме може изразити комплексније атрибуте попут хармоније, разликовања музичких инструмената и слично. Слика 4.1 приказује описане компоненте конволутивних неуронских мрежа коришћених у обради звука.

Поред поделе по врсти улаза, постоји и подела конволутивних мрежа за класификацију звука и на моделе нивоа инстанце (енгл. *instance level*) и моделе нивоа песама (*song-level*). Оба типа модела се користе за класификацију целих песама. Разлика је што први тип модела не прима целу песму на улазу већ прима само неки њен сегмент (нпр. секвенцу од неколико секунди) на основу чега врши класификацију песме, док други тип модела прима целу песму. Предност првог типа модела је у томе што ће величина улаза бити фиксирана (нпр. три секунде) што модел чини мање комплексним у погледу израчунавања, а типично оваквав приступ даје нешто боље перформансе у смислу класификације. Како би се додатно повећала робусност првог типа модела, у пракси се типично песма за класификацију подели у сегменте, сваки сегмент се анализира моделом, а затим се већинским гласањем одређује крајњи резултат. Предност модела нивоа песама се огледа у томе што они не

захтевају ни препроцесирање нити имплементацију додатне логике након што изврше предвиђање. У наставку следи опис неколико модела који (у тренутку писања овог рада) пријављују најбоље резултате. Сви модели су инстанцног типа, неки постоје само у форми предњег дела за спектрограме, неки за аудио запис, а неки постоје у обе варијанте.

Потпуно конволутивна мрежа – ФЦН

Потпуно конволутивна мрежа (енгл. *Fully convolutional network, FCN*) [4] је једна од првих архитектура конволутивних мрежа коришћених за аутоматско означавање песама. Скупови података над којима је овај модел обучаван и евалуиран су Магна таг тјун и Скуп милион песама. Овај модел припада породици модела на нивоу песама. Улаз у ову мрежу је песма у трајању од 29 секунди (што је типична дужина песама у популарним скуповима података) трансформисана у мелодијски спектрограм. Како су димензија улазне песме и фреквенција узорковања увек исти, тако свака мапа атрибута увек има исте димензије, па није неопходно примењивати никакве стратегије поравнавања излаза већ ће последњи конволутивни слој ФНЦ увек имати 50 неурона што и одговара броју категорија у скуповима над којима је модел обучаван. За разлику од класичних конволутивних мрежа које поседују неколико потпуно повезаних слојева на самом крају којима се врши класификација, овај модел класификацију ради применом одговарајуће активационе функције (сигмоидне или функције меког максимума) директно над мапом атрибута. Из недостатка потпуно повезаних слојева и следи назив овог модела.

У зависности од броја конволутивних слојева у моделу, аутори су евалуирали неколико FCN_K за $K \in \{3, 4, 5, 6, 7\}$, где број K означава управо количину конволутивних слојева. Табела 4.2 приказује пријављене мере квалитета неколико различитих ФЦН модела мерене површином испод РОК криве.

Конволутивна мрежа на нивоу узорка

Конволутивна мрежа на нивоу узорка (енгл. *Sample-level CNN*) [15] је једноставна конволутивна неуронска мрежа која на улазу прима нетрансформисан аудио сигнал и на основу узорка тог сигнала врши жанровску класификацију сигнала. У основи, овај модел је инспирисан класификационим моделом ВГГ (енгл. *VGG*) [25] из области рачунарске визије. Овај модел се састоји од

Модел	Скуп података	
	MTT	СМП
FCN_3	0,852	0,786
FCN_4	**0,894**	0,808
FCN_5	0,890	0,848
FCN_6	-	**0,851**
FCN_7	-	0,845

Табела 4.2: Пријављени резултати ФЦН модела [4]

једнодимензионих конволутивних слојева (након сваког конволутивног слоја је постављен слој агрегације максимумом) при чему је последњи слој мреже један потпуно повезан слој са сигмоидном активацијом. Како ради над узорком сигнала, припада класи инстанцних модела.

Аутори овог модела су вршили разне експерименте углавном варирајући величину конволутивних филтера, дужину улазне секвенце и количину конволутивних слојева. Тако су увели номенклатуру модела – m^n при чему m представља величину (једнодимензионих) конволутивних филтера док n означава количину конволутивних слојева. Пример једног модела ознаке 3^9 који на улазу прима секвенцу трајања $2678ms$ узорковану фреквенцијом $22050Hz$ је приказан табелом 4.3. Независно од ознаке модела, у свим конволутивним слојевима је за функцију активације коришћена изломљена линеарна јединица. Поред тога су излази свих слојева нормализовани на нивоу групе инстанци коришћене у једној итерацији пропагације уназад (енгл. *batch normalization*) [12].

Аутори су избор модела вршили користећи Магна таг тјун скуп података, док су евалуирали и поредили са другим моделима над скуповима Магна таг тјун и Скупом милион песама. Метрика која је у оригиналном раду коришћена и за избор мета параметара и за крајњу евалуацију модела је површина испод РОК криве. Табела 4.4 показује остварене метрике при различитим изборима метапараметара m и n и величине улазног сигнала, док табела 4.5 приказује пријављене резултате у поређењу са другим моделима који су били актуелни до 2017. године када је овај рад објављен.

Један од важнијих доприноса коју је овај модел донео је чињеница да улаз у мрежу није процесиран (нпр. сигнал није претваран у спектрограме или нешто слично) већ је изворни запис звука оно над чиме је модел обучаван.

слој	померај	излаз	# параметара
конв 3-128	3	19683×128	514
норм	-	19683×128	
конв 3-128	1	19683×128	
макс 3	3	6561×128	49536
норм	-	6561×128	
конв 3-128	1	6561×128	
макс 3	3	2187×128	49536
норм	-	2187×128	
конв 3-128	1	2187×256	
макс 3	3	729×256	99072
норм	-	729×256	
конв 3-128	1	729×256	
макс 3	3	243×256	197376
норм	-	243×256	
конв 3-128	1	243×256	
макс 3	3	81×256	197376
норм	-	81×256	
конв 3-128	1	81×256	
макс 3	3	27×256	197376
норм	-	27×256	
конв 3-128	1	27×256	
макс 3	3	9×256	197376
норм	-	9×256	
конв 3-128	1	9×256	
макс 3	3	3×256	197376
норм	-	3×256	
конв 3-128	1	3×512	
макс 3	3	1×512	394752
норм	-	1×512	
конв 3-128	1	1×512	
норм	-	1×512	263680
изостављање 0.5	3	1×512	
потпуно пов.	-	50	25650
Укупан број параметара			1.9×10^6

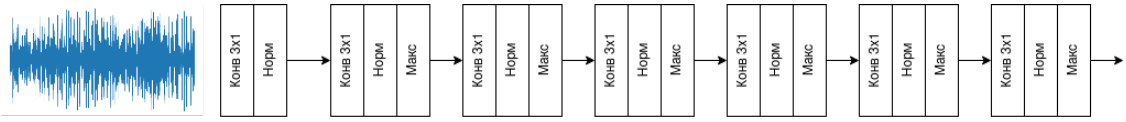
Табела 4.3: Модел конволутивне мреже на нивоу узорка ознаке 3^9 са улазом трајања 2678ms [15]

m^n	трајање (ms)	РОК
2^{10}	743	0,9011
2^{14}	1486	0,9040
3^8	893	0,9039
3^9	**2678**	**0,9055**
4^6	743	0,9021
4^7	2972	0,9026
5^5	709	0,9024
5^6	3543	0,9041

Табела 4.4: Избор модела на основу метапараметара [15]

Модел	Скуп података	
	МТТ	СМП
Постојани ЦНН	0,9013	-
1Д ЦНН	0,8487	-
2Д ЦНН	0,894	0,851
ЦРНН	-	0,862
Предложени модел	**0,9055**	**0,8812**

Табела 4.5: Пријављени резултати модела Конволутивна мрежа на нивоу узорка 3^9 [15]



Слика 4.2: Предњи део Мјузишн модела (аудио) [22]

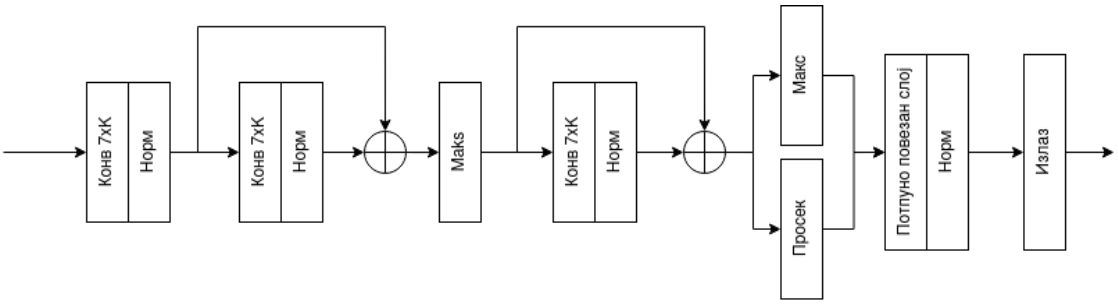
Мјузишн

Модел Мјузишн (енгл. *MusiCNN*) [22] припада групи инстанцних модела и обучаван је над сегментима од по три секунде. Скупови података коришћени за обучавање Мјузишн модела су Магна таг тјун, Скуп милион песама, као и приватни скуп података којим аутори располажу обима 1.2 милиона песама. Рад је објављен од стране истраживача компаније Пандора, једне од већих компанија у области емитовања мултимедијалног садржаја, што објашњава и обим и квалитет приватног скупа над којим су обучавали модел. Оригинално су предложене две верзије овог модела: верзија са предњим делом који ради над аудио сигналом и верзија са предњим делом која ради над (мелодијским) спектрограмима.

Прва верзија је конструисана користећи једнодимензионе конволутивне слојеве. Предњи део ове мреже је инспирисан претходно описаном конволутивном мрежом на нивоу узорка. Главна идеја иза овог приступа је била конструисање конволутивних филтера не укључујући доменско знање. Конволутивни филтери предњег дела би као и код претходног модела сами научили да издвајају неопходне карактеристике звука. На слици 4.2 је приказана схема предњег дела мреже.

Верзија модела која ради над спектрограмима укључује доменско знање приликом дизајна конволутивних филтера. Спектрограми (било стандардни или мелодијски) имају своју просторну интерпретацију – хоризонтална оса одговара временској компоненти сигнала док вертикална оса одговара фреквенцијама у сигналу. Одатле следи да конволутивни филтери који су вертикалних пропорција (којима је вертикална компонента већа) одговарају детектовању боје звука, док филтери веће хоризонталне димензије лакше обухватају временске зависности у сигналу [21]. У оригиналном раду, аутори су предложили вертикалне филтере чија је ширина фиксирана на седам јединица временске скале (што је ефективно седам пиксела на спектрограму) док им је висина дефинисана метапаремтром чије вредности могу ићи и до 80% укупне висине спектрограма. Хоризонтални филтери су углавном само транспоновани тако да им је хоризонтална компонента веће дужине. Сама конструкција мреже оваквим филтерима је такође атипична где се слојеви мреже не надовезују један на други продубљујући мрежу. Уместо тога, конструисано је шест слојева користећи филтере вертикалних пропорција и четири слоја који садрже филтере хоризонталних пропорција, при чему сваки слој на улазу прима улазни мелодијски спектрограм. Након рачунања атрибута, излази свих десет слојева се конкатенирају у једну финалну мапу атрибута која представља излаз предњег дела мреже. Користећи вертикалне филтере након којих се налази слој агрегације максимумом, овај део мреже успева да постигне инваријантност у односу на висину тона, те тиме изолујући саму боју звука [23].

Задњи део мреже је идентичан у оба случаја и сграђен је користећи дводимензионе конволутивне филтере. Како је ово већ дубљи део мреже, аутори су предложили коришћење прескакајућих веза (енгл. *skip-connections, residual connections*) те се на тај начин стабилизовало обучавање мреже [16]. Испробавањем разних дубина задњег дела мреже се закључило да се перформансе модела не поправљају драстично тако да је финална верзија садржала три конволутивна слоја. Пре изравнавања мапе атрибута зарад постављања потпуно повезаних слојева, мапе атрибута се независно агрегирају слојевима агрегације просеком и максимумом, али филтерима варијабилних димензија. На овај начин се омогућило да потпуно повезани слојеви буду увек истих димензија независно од величине улазног сигнала. Скица архитектуре задњег дела се може видети на слици 4.3.



Слика 4.3: Задњи део Мјузишн модела (параметар K зависи од излазне мапе атрибута предњег дела) [22]

Модел	Приватан скуп		СМП		МТТ	
	РОК	ПО	РОК	ПО	РОК	ПО
ЦНН – Узорак	-	-	0,8812	-	0,8865	0,3438
Мјузишн (аудио)	0,9250	0,6120	0,8741	0,2853	0,8905	0,3492
Мјузишн (спектрограм)	0,9217	0,5929	0,8875	0,3124	0,9040	0,3811

Табела 4.6: Пријављене мере квалитета модела Мјузишн [22]

Поређење обе верзије модела са конволутивном мрежом за рад са узорком се може видети у табели 4.6 где су излистани резултати поређења над сва три коришћена скупа мерена површином испод РОК криве и површином испод прецизност-одзивност криве. Из тих резултата се јасно види да је за мање скупове података доменско знање допринело квалитету модела, док се на већим скуповима модел који директно ради над аудио сигналом понаша боље.

4.3 Аугментације података

Како би се надоместио недостатак података и поред тога повећала робусност модела, примењено је неколико аугментација над подацима за обучавање. Примењивање аугментација није детерминистички процес већ је приликом обуке модела у фази учитавања података свака аугментација примењена са одређеном вероватноћом користећи библиотеку „Torchaudio Augmentations” [26].

Случајни исечак

Модели који су коришћени припадају класи модела нивоа инстанце и обучавани су на сегментима од три секунде. Како су све инстанце у овом подскупу трајања тридесет секунди, основна аугментација је одабир насумичног одсечка песме (енгл. *random resized crop*). Ова аугментација је уједно и једина аугментација која ће се применити над свим инстанцама, са том разликом што ће свако њено примењивање произвести другачији исечак, односно исећи ће други део песме.

Појачавање сигнала

Како би модели постали робусни на разне нивое гласноће сигнала неопходно је да сâм процес обучавања модела наиђе на разне нивое гласноће звука. Гласноћа звука се контролише појачавањем сигнала (енгл. *gain*), односно множењем сигнала реалним бројем. Слободно говорећи, промена гласноће сигнала *је аналогна* униформној промени осветљења фотографије у рачунарској визији и честа је техника аугментације звучног сигнала. Појачавање сигнала је параметризовано вредностима g_{min} и g_{max} што одговара минималном и максималном појачавању тим редом, при чему се сигнал заправо појачава насумичном вредношћу из интервала $[g_{min}, g_{max}]$. Јединица у којој се појачавање сигнала изражава су децибели.

Промена знака

Промена знака (енгл. *polarity inversion*) представља специјалан случај појачавања сигнала где се сигнал уместо произвољним реалним бројем множи негативном јединицом. Овим ће амплитуде сигнала (којима се добија звук) остати исте, али ће сигнал бити промењеног знака. Када се неколико сигнала спајају у један сигнал, промена знака појединачних сигнала може утицати на крајњи резултат. Због тога је то техника која се често користи приликом продукције музике, конкретно током мешања звука (енгл. *sound mixing*) како би се избегла нежељена својства која природно настају мешањем (попут поништавања сигнала или појаве дисторзије). Промена знака појединачног сигнала нема никаквог утицаја на то како он сâм звучи што је чини погодном техником аугментације података.

Шум

Додавање случајног шума помаже моделу да постане отпоран на разне проблеме који се појављују у сигналу, попут природно индукованог шума услед снимања у неадекватним условима, затим дисторзије сигнала и слично. Случајни шум се додаје тако што се сигнал сабира случајним вектором на тај начин да се постигне жељени однос сигнала и шума (енгл. *signal to noise ratio – SNR*). Нека су x и n вектори исте дужине који представљају улазни сигнал и случајни вектор шума, тим редом и нека је SNR жељени однос сигнала и шума изражен у децибелима. Додавање шума се врши на следећи начин [32]:

$$y = x + an \tag{4.1}$$

$$a = \sqrt{\frac{\|x\|_2^2}{\|n\|_2^2} \cdot 10^{-\frac{SNR}{10}}} \tag{4.2}$$

Реверберација

Када се слушалац налази у центру просторије добрих акустичних својстава, тада звук који долази са позорнице не долази искључиво директно са бине већ, како се звук простире по свим правцима, одбијањем од зидова просторије такође долази до слушаоца. Обзиром да звук који долази из ових алтернативних праваца прелази веће растојање, самим тиме долази са благим закашњењем у односу на примарни звук који долази директно са бине. Поред тога, услед рефлексије долази и до губитка енергије те до слушаоца овај звук стиже нешто тиши. Овај ефекат се назива реверберација звука и она чини да је звук перципиран пуније и природније.

Музичка продукција данас у великој мери примењује дигиталну реверберацију сигнала. Музика се снима у глувим собама како се ниједан ефекат не би природно појавио у снимљеном звуку, затим се приликом монтаже додају разни ефекти. Аутори можда у већој или у мањој мери примењују реверберацију звука, али је овај ефекат свакако свеприсутан. Како сам ефекат нема утицаја на жанр, ово је погодан начин аугментовања. Типично, имплементације дигиталне реверберације су параметризоване интензитетом реверберације и величином собе.

Корекција висине тона

Повремено се дешава да аутори напишу песму у једном музичком кључу, а уживо изведу у другом. Такође, аутори током интерпретације музике других аутора адаптирају песме свом сензибилитету те често промене кључ. Тиме што је цела песма повишена за неколико тонова или полутонова (или чак само благо измештена из тоналитета) се не мења жанр те песме. Неке жанрове попут би-бапа чак одликују честе промене кључа током једне песме[1].

[1]Би-бап је стил у џезу који је био доминантан четрдесетих година двадесетог века. Карактеристичан је по брзом темпу, комплексним ритмовима, честим променама кључа и технички захтевним композицијама.

Глава 5

Експерименти и резултати

Истакнути модели над којима су вршени експерименти су Мјузишн и Конволутивна мрежа на нивоу узорка. Како би поређење било релевантно и како би модели постигли што боље перформансе, коришћен је скуп CMA. Пошто је обучавање модела рачунарски сувише захтевно на великим скуповима података, за обучавање модела је коришћен подскуп скупа CMA – *мали CMA*. Скуп CMA поседује хијерархијску таксономију класа, па је сваки музички жанр који се појављује у овом скупу или примарни (основни) жанр или хијерархијски директно или индиректно припада неком примарном жанру. У овом подскупу је осам хиљада инстанци равномерно распоређено по примарним жанровима те је класификација овако балансираних жанрова утолико лакша. Подскуп *мали CMA* садржи инстанце које припадају тачно једном од осам примарних жанрова:

- Интернационална музика – (скр. *ИМ*)
- Експериментална музика – (скр. *ЕМ*)
- Поп
- Рок
- Електронска музика (скр. *ЕДМ*)
- Фолк
- Инструментална музика (скр. *ЕДМ*)
- Хип–хоп

5.1 Експерименти

Међу многим експериментима, селектована су три главна експеримента за приказ у овом раду. Конкретно, обучена су следећа три модела:

- Конволутивна мрежа на нивоу узорка (конфигурације 3^9)
- Мјузишн модел за рад над аудио сигналом
- Мјузишн модел за рад над спектрограмима

Подела података на скупове за обучавање, избор модела и евалуацију је конзистентна за сва три експеримента и извршена је према предложеној подели аутора CMA скупа. Како су оба модела заснована на инстанцама, дужина исечка за коју су дефинисани модели је три секунде при фреквенцији узорковања $44100Hz$. Избор аугментација је такође истоветан како би експерименти били упоредиви, а конфигурација параметара је описана табелом 5.1.

За избор метапараметара је коришћен скуп за избор модела пре покретања наведених експеримената, а током наведених експеримената се овај скуп користио само за детектовање преприлагођавања модела. Након обучавања, сви модели су евалуирани користећи скуп за евалуацију модела који током фаза обучавања и избора модела није коришћен. Скуп за евалуацију модела је такав да не постоји преклапање аутора музике са скуповима за обучавање и избор модела, тако да је музика која се користила за евалуацију репрезентативан узорак нове и непознате музике.

Конволутивна мрежа на нивоу узорка

Основни модел обучаван за класификацију музике на жанрове је конволутивна мрежа на нивоу узорка конфигурације 3^9. Модел је обучаван из шест етапа при чему је свака етапа вршила обуку модела кроз сто епоха. Даље обучавање модела је водило преприлагођавању па је шеста етапа узета као финална. За оптимизацију је коришћен Радам оптимизатор са иницијалним кораком учења 0.003 [17]. Како корак учења код Радам оптимизатора није константан кроз епохе, једно обучавање модела у трајању од 600 епоха није исто као шест етапа обучавања у трајању од по сто епоха. Табела 5.2 приказује резултате финалног модела мерене на скупу за евалуацију.

Случајни исечак	Обим исечка: $3 \cdot 44100$
Промена знака	Вероватноћа: 0.5
Шум	Вероватноћа: 0.3 Минимални SNR: $5 \cdot 10^{-3}$ Максимални SNR: $2 \cdot 10^{-2}$
Појачавање сигнала	Вероватноћа: 0.4 Минимално појачавање: $-20db$ Максимално појачавање: $-1db$
Реверберација	Вероватноћа: 0.25 Минимална реверберација: 50 Максимална реверберација: 91 Минимална величина собе: 80 Максимална величина собе: 101
Корекција висине	Вероватноћа: 0.3 Минимално корекција: -7 Максимална корекција: 7

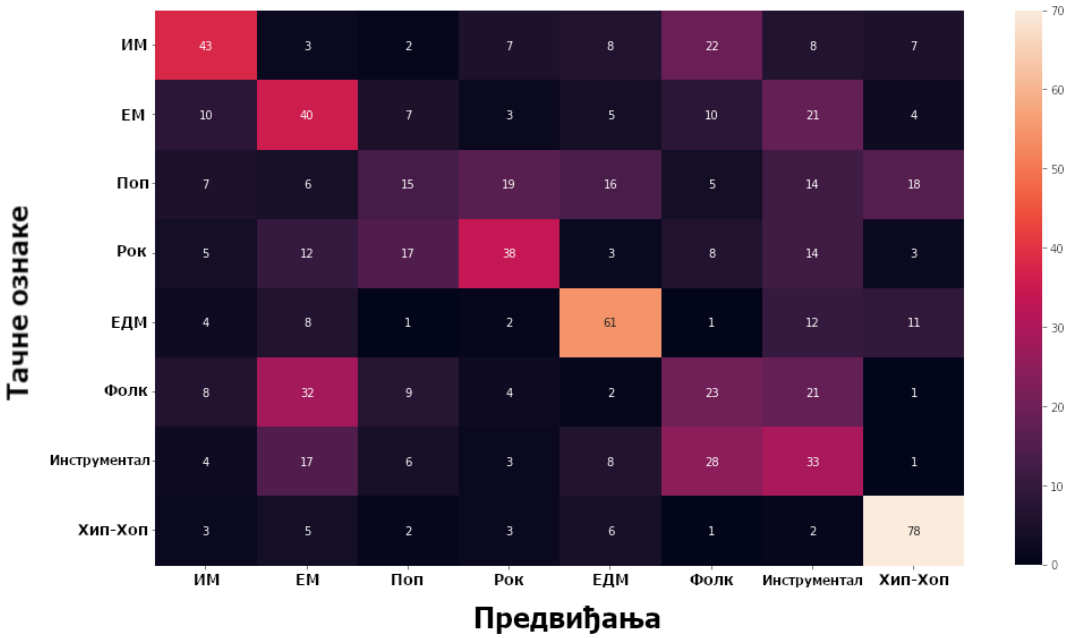
Табела 5.1: Конфигурација параметара аугментације података

Жанр	Прецизност	Одзив
ИМ	0.37	0.54
ЕМ	0.25	0.16
Поп	0.33	0.01
Рок	0.55	0.59
ЕДМ	0.33	0.50
Фолк	0.57	0.64
Инструментал	0.39	0.33
Хип-Хоп	0.53	0.70
Просек	0.42	0.43
Тачност: 0.43 $ROC - AUC$: 0.79 Обим узорка: 800		

Табела 5.2: Постигнуте метрике конволутивне мреже на нивоу узорка

Слика 5.1 показује матрицу конфузије на којој се детаљније види понашање модела. Из матрице се види да је модел извесне жанрове лакше савладао од других. На пример, највећи део фолк песама је сврстао у категорију експерименталне музике док је са друге стране песме електронске музике већински класификовао исправно. Пријављене $ROC - AUC$ метрике су нешто веће од постигнутих, што је очекивано обзиром на незанемарљиву разлику у обиму

података.



Слика 5.1: Матрица конфузије конволутивне мреже на нивоу узорка

Мјузишн (аудио верзија)

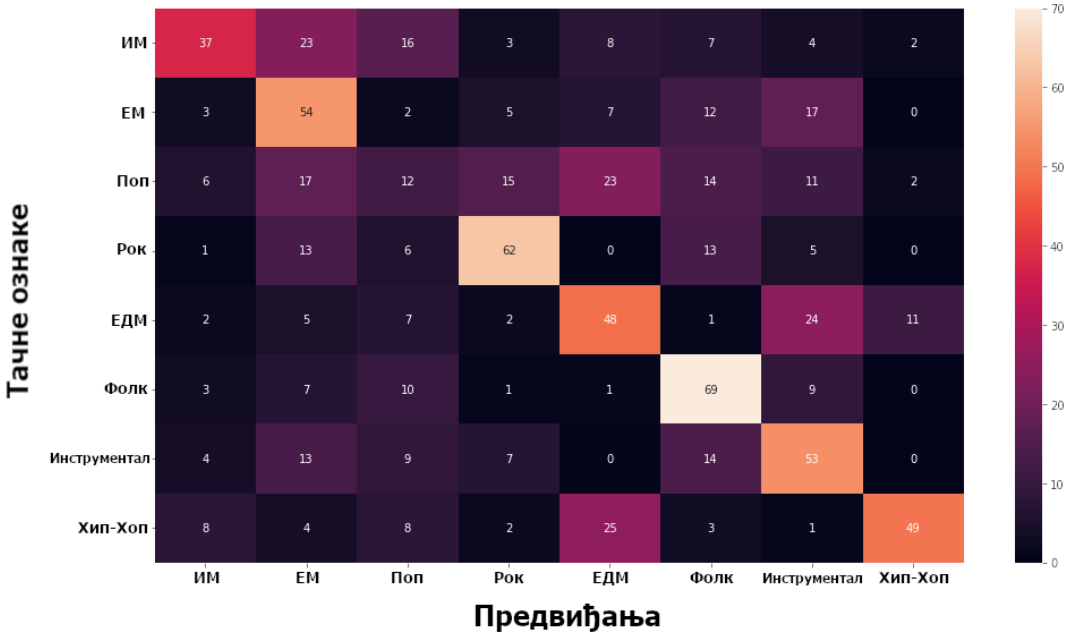
Модел Мјузишн који ради над изворним аудио датотекама је дозволио обимније обучавање у односу на конволутивну мрежу на нивоу узорка. Метапараметри који се тичу оптимизације су остали идентични – оптимизација је извршена Радам оптимизатором чији је иницијални корак учења био 0.003 и трајање етапе је такође било сто епоха. За разлику од претходног модела, Мјузишн је прошао кроз 16 етапа. Даље обучавање модела није донело никаквог прогреса ни на скупу за обучавање ни на скупу за избор модела.

Табела 5.3 показује добијене резултате на скупу за евалуацију. Све агрегиране метрике попут укупне тачности модела, просечне прецизности и просечног одзива су боље у односу на претходни модел. Ипак, неке жанрове као што је на пример поп, овај модел класификује лошије. Битно је истаћи да је Мјузишн био смелији да инстанцу класификује жанром поп што се огледа у већем одзиву овог жанра, а приликом повећања одзива је очекиван пад прецизности. Код овог модела је разлика између пријављених резултата на другим скуповима и добијених резултата на CMA скупу мања него што је то

Жанр	Прецизност	Одзив
ИМ	0.58	0.37
ЕМ	0.40	0.54
Поп	0.17	0.12
Рок	0.64	0.62
ЕДМ	0.43	0.48
Фолк	0.52	0.69
Инструментал	0.43	0.53
Хип-Хоп	0.77	0.49
Просек	0.49	0.48
Тачност: 0.48 $ROC - AUC$: 0.81 Обим узорка: 800		

Табела 5.3: Постигнуте метрике модела Мјузишн (аудио)

случај са претходним моделом. Матрица конфузије приказана сликом 5.2 даје мало бољу слику где су се предвиђања за поп музику распршиле, односно са којим жанровима је модел мешао друге жанрове.



Слика 5.2: Матрица конфузије модела Мјузишн

Мјузишн (спектрограм)

Модел Мјузишн који ради над спектрограмима је досегао до три етапе (при чему свака етапа такође износи сто епоха) пре него што је примећено преприлагођавање. Оптимизација Радам оптимизатором без обзира на избор метапараметара није довела до обучавања модела па је код овог експеримента примењен стандардни градијентни спуст. Модел је обучаван фиксним кораком учења 0.005 кроз 300 епоха пре него што је уочено преприлагођавање.

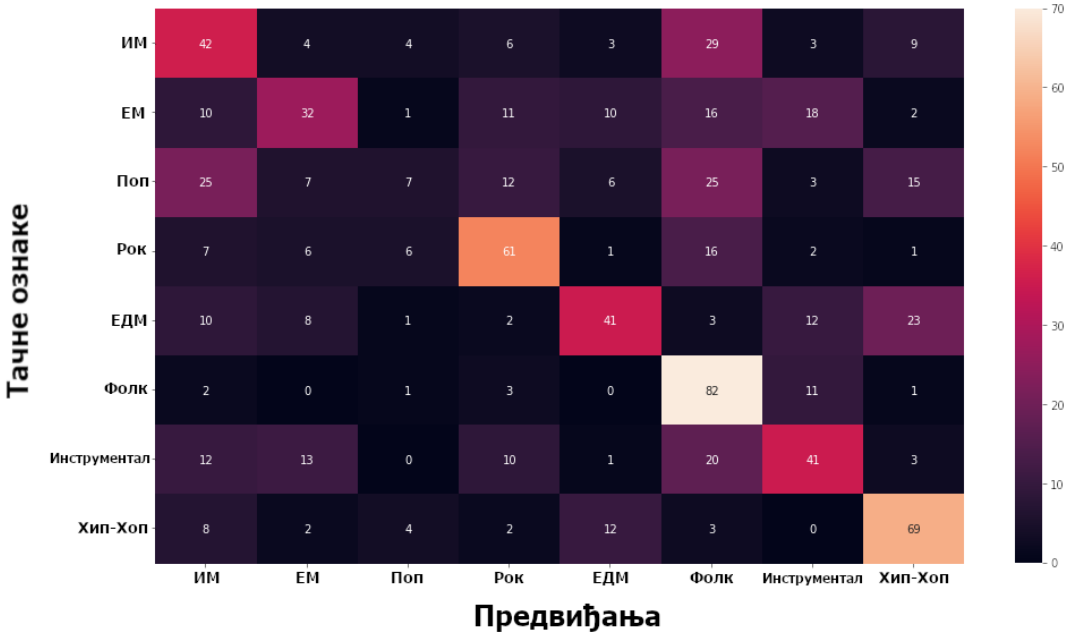
Жанр	Прецизност	Одзив
ИМ	0.36	0.42
ЕМ	0.44	0.32
Поп	0.29	0.07
Рок	0.57	0.61
ЕДМ	0.55	0.41
Фолк	0.42	0.82
Инструментал	0.46	0.41
Хип-Хоп	0.56	0.69
Просек	0.46	0.47
Тачност: 0.47 $ROC - AUC$: 0.82 Обим узорка: 800		

Табела 5.4: Постигнуте метрике модела Мјузишн (спектрограм)

Резултати постигнути овим моделом су махом слични као код Мјузишн модела који директно користи звучни сигнал, с том разликом да је $ROC - AUC$ метрика за нијансу боља код верзије модела која ради са спектрограмом. Овакав однос та два модела је описан и у оригиналном раду. Аутори су пронашли да је за мање скупове података боље резултате дала верзија модела која укључује доменско знање тј. модел који користи спектрограме над којима примењује хоризонталне и вертикалне филтере. Мере квалитета овог модела су приказане у табели 5.4.

Слика 5.3 приказује матрицу конфузије овог модела. Као и код претходних модела, поп је био један од тежих жанрова за класификацију. За разлику од претходних модела, овај модел се доста боље показао при класификацији песама експерименталне музике. Обзиром да је модел обучаван кроз само 300 епоха, очекивано је било да ће се десити пристрасност модела ка једном жанру. У овом експерименту је то био фолк, што се види и по томе да је одзив

те класе био чак 0.82 што је највећи постигнут одзив међу свим моделима и свим жанровима до сада. Пошто су класе балансиране у овом скупу, то што је фолк испао као жанр који је модел најчешће предвиђао је највероватније случајно. Дохватање података се вршило случајним избором без враћања па је могуће да су последње итерације последње епохе имале несразмерно већи број фолк песама. Како би се ово детаљније утврдило неопходно је извршити још експеримената.



Слика 5.3: Матрица конфузије модела Мјузишн (спектрограм)

5.2 Поређење резултата

Пред почетак обучавања ових модела параметри модела су иницијализовани насумично. Ово није увек случај, често се у пракси користе већ обучени модели те се обука новим задатком само надовезује на претходну обуку у смислу да почетна тачка градијентног спуста није насумична већ се за почетну тачку узима крајња тачка претходне обуке. Овај процес се назива преношење знања (енгл. *transfer learning*) и представља лак и ефикасан начин обучавања модела.

Три модела са којима су вршени експерименти су показали добре резултате обзиром да су обучавани над веома малим скупом података и то „од нуле". Табела 5.5 приказује резултате свих модела при чему су за сваки жанр истакнути модели који су постигли највећу прецизност и највећи одзив независно. Резултати су веома блиски једни другима, али поред тога је приметно да неки модели лакше класификују неке жанрове од других. Овај феномен је, као што је претходно речено, врло вероватно последица насумичног редоследа песама у фази обуке модела.

Жанр	ЦНН – Узорак		Мјузишн (аудио)		Мјузишн (спектрограм)	
	Прецизност	Одзив	Прецизност	Одзив	Прецизност	Одзив
ИМ	0.37	**0.54**	**0.58**	0.37	0.36	0.42
ЕМ	0.25	0.16	0.40	**0.54**	**0.44**	0.32
Поп	**0.33**	0.01	0.17	**0.12**	0.29	0.07
Рок	0.55	0.59	**0.64**	**0.62**	0.57	0.61
ЕДМ	0.33	**0.50**	0.43	0.48	**0.55**	0.41
Фолк	**0.57**	0.64	0.52	0.69	0.42	**0.82**
Инструментал	0.39	0.33	0.43	**0.53**	**0.46**	0.41
Хип-Хоп	0.53	**0.70**	**0.77**	0.49	0.56	0.69
Просек	0.42	0.43	**0.49**	**0.48**	0.46	0.47
Тачност		0.43		**0.48**		0.47
$ROC - AUC$		0.79		0.81		**0.82**

Табела 5.5: Поређење свих модела

Глава 6

Закључак

Анализирањем резултата је примећено да су се свим моделима неки од жанрова попут поп музике и експерименталне музике показали тежим за класификацију. Ово и није изненађујуће обзиром на то да оба ова жанра представљају кровни термин за велики број поджанрова, те ове две класе имају велику унутрашњу варијансу. Поред тога, сам жанр као такав је слабо дефинисан концепт и веома често је класификација песме у неки жанр субјективна ствар. Класе чија је унутрашња варијанса мала, тј. они жанрови који се објективно могу сматрати препознатљивим као што су хип-хоп или електронска музика су лакше били класификовани од стране сва три модела. Ово указује на то да су модели успели да опишу унутрашњу музичку структуру попут хармоније, ритма, темпа и да то вежу за одговарајући жанр.

Највећи технички изазови су недостатак података, односно недостатак ресурса који би омогућили обучавање модела над већим обимом података. Примера ради, сто епоха обучавања ових модела је трајало недељу дана по моделу користећи једну графичку картицу марке Nvidia ознаке 1070ti. За обучавање већих модела или за обучавање на већем скупу података је неопходан перформантнији хардвер. Таква обучавања су знатно скупља, делом због саме цене хардвера, а делом и због утрошене електричне енергије што представља велику препреку самосталног истраживачког процеса.

У даљем раду би требало размотрити модерне архитектуре које су постале популарне у другим гранама машинског учења као и обучавање модела над већим скупом података. Такође, поред класификације музике на примарне жанрове, било би интересантно видети како би модели савладали задатак класификације на изведене жанрове. Тиме би се решио проблем велике вари-

јансе унутар жанрова пошто би се кровни жанрови који имају велику варијансу разбили на одговарајуће поджанрове. Ипак, изазов у овом приступу је лоша балансираност класа јер су неки изведени жанрови заступљени у свега неколико песама док су неки други заступљени у неколико хиљада песама.

Библиографија

[1] About spotify, Jul 2022.

[2] Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, and Paul Lamere. The million song dataset. pages 591–596, 01 2011.

[3] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. In *International Conference on Machine Learning*, 2019.

[4] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks, 2016.

[5] Nicholas J. Conard, Maria Malina, and Susanne C. Münzel. New flutes document the earliest musical tradition in southwestern germany. *Nature*, 460(7256):737–740, Aug 2009.

[6] David Curry. Music streaming app revenue and usage statistics (2022), Sep 2022.

[7] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis, 2016.

[8] Franco Fabbri. A theory of musical genres. two applications. 1982.

[9] Gustav Theodor Fechner. *Psychophysiology.* v.1. Breitkopf und Härtel, Leipzig, 1860.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[11] Thomas Higham, Laura Basell, Roger Jacobi, Rachel Wood, Christopher Bronk Ramsey, and Nicholas J. Conard. Testing models

for the beginnings of the aurignacian and the advent of figurative art and music: The radiocarbon chronology of geißenklösterle. *Journal of Human Evolution*, 62(6):664–676, 2012.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[14] Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. S. Downie. Evaluation of algorithms using games: The case of music tagging. In *International Society for Music Information Retrieval Conference*, 2009.

[15] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms, 2017.

[16] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets, 2018.

[17] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond, 2021.

[18] Dr. Mladen Nikolić. Mašinsko učenje, 2019.

[19] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley series in electrical engineering. Addison-Wesley Publishing Company, 1987.

[20] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project, 2004.

[21] Jordi Pons, Thomas Lidy, and Xavier Serra. Experimenting with musically motivated convolutional neural networks. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2016.

[22] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale, 2018.

[23] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra. Timbre analysis of music audio signals with convolutional neural networks, 2017.

[24] Rajat Raina, Anand Madhavan, and Andrew Ng. Large-scale deep unsupervised learning using graphics processors. *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, 382:110, 01 2009.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[26] Janne Spijkervet. Spijkervet/torchaudio-augmentations, 2021.

[27] Bob L. Sturm. The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, 43(2):147–172, apr 2014.

[28] Bob L. Sturm. A survey of evaluation in music genre recognition. In Andreas Nürnberger, Sebastian Stober, Birger Larsen, and Marcin Detyniecki, editors, *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, pages 29–66, Cham, 2014. Springer International Publishing.

[29] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[30] Avery Wang et al. An industrial strength audio search algorithm. In *Ismir*, volume 2003, pages 7–13. Washington, DC, 2003.

[31] Minz Won, Keunwoo Choi, and Xavier Serra. Semi-supervised music tagging transformer, 2021.

[32] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy,

Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*, 2021.

Биографија аутора

Урош је рођен 13. септембра 1989. године у Смедереву. Као мали се сели у Опово, где завршава првих пет разреда основне школе Доситеј Обрадовић. 2001. године се сели у Панчево где завршава основну школу Исидора Секулић. Још од малена је показивао склоност ка инжењерији, те средњу машинску школу „Панчево" уписује 2004. године. По завршетку средње школе и двогодишње паузе уписује Математички факултет у Београду. Самостално припрема пријемни испит и 2010. године уписује смер Информатика на Математичком факултету. Током друге године факултета одлази на конференцију ЕТРАН где у секцији за вештачку интелигенцију претзентује групни рад Логички калкулатор који имплементира синтаксно окружење за трансформације формула исказне и предикатске логике и испитивање ваљаности ових формула, посредно, методом резолуције. Након друге године факултета почиње да ради као програмер и свој примарни фокус са студија пребацује на индустрију. Пар година касније се сели у Београд, пребацује на нову акредитацију смера Информатика, полаже преостале испите и тиме завршава основне студије 2018. године. Током основних студија се заинтересовао за теорију машинског учења, због чега се 2016. године пријављује и одлази да похађа семинар посвећен машинском учењу ПСИМЛ. Поучен лепим искуством, 2018. године одлази у Клуж (Румунија) на летњу школу машинског учења ТМЛСС у организацији компаније Дип Мајнд. Исте године се запошљава у фирми Еверсин на позицији истраживача из области рачунарске визије. Радећи на овој позицији и решавајући разне практичне проблеме бива унапређен и почиње да води истраживачки тим. Услед потреба посла, креће да ради на споредном пројекту – развој дистрибуиране платформе за анотирање података. Крајем 2020. године напушта истраживачки посао и преузима у потпуности развој платформе. Као предавач и ментор се прикључује ПСИМЛ семинару 2019. године на којем држи предавања из области конволутивних неуронских мрежа и детекције објеката, а поред предавања води радионице логистичке регресије и конволутивних неуронских мрежа. Музика је једна од његових највећих страсти. Љубав ка музици гаји од малена и она је од тада обележила велики део његовог живота. Инспирисан овиме одлучује да се у мастер раду бави музиком. Тако и бира тему „Аутоматска жанровска класификација песама техникама дубоког учења" ком заинтригирано посвећује две године истраживачког рада.