Eötvös Loránd Tudományegyetem

Informatikai Kar

Programozáselmélet és Szoftvertechnológiai

Tanszék

Automatikus zenei hangszerfelismerés többszólamú zenében mély neuronhálók segítségével

Témavezető:
Gombos Gergő
Adjunktus, PhD

Szerző:
Hamrák János
programtervező informatikus MSc

Budapest, 2020

Az eredeti szakdolgozati / diplomamunka témabjelentő helye.

Tartalomjegyzék

1. fejezet

Bevezetés

Napjainkban a zenéhez legkönnyebben digitális formában, a világhálón keresztül férhetünk hozzá. Néhány kattintással olyan zenei tartalomszolgáltatókat érhetünk el, melyek széleskörű adatbázissal rendelkeznek. A folyamatosan bővülő adatmennyiség ellenére ezeknek az adatbázisoknak átláthatónak és könnyen kezelhetőnek kell maradniuk, hogy a felhasználókat a kívánt módon tudják kiszolgálni. Ennek érdekében nap mint nap új megoldások születnek zenei információk automatikus kinyerése és feldolgozása céljából. Ezek teszik lehetővé a digitálisan tárolt zenék körében például az osztályozást vagy keresést.

A zenei információk kinyerésének tudományába (Music Information Retrieval, a továbbiakban: "MIR") tartozik a továbbiakban taglalt probléma, az automatikus hangszerfelismerés. Ez egy osztályozási feladat. Célja, hogy a meglévő digitális hanganyag alapján az adott zenéről megállapítsuk, hogy milyen hangszerek szólalnak meg benne. Ezt az információt több célra is fel tudjuk használni, például:

- Későbbi feldolgozásra, MIR feladatok inputjaként
- Statisztikák készítésére
- Adatbázisban való keresés szűrőfeltételeként
- Egy ajánlórendszer részeként, ahol az aktuális zeneszámot követően például egy hangszerelésében hasonló számot szeretnék ajánlani a felhasználónak.

Az automatikus hangszerfelismerés feladatot több aspektusból lehet megközelíteni, például a bemeneti adatok jellege, reprezentációja, a megvalósított architektúra,

az osztályozás módszere, vagy az osztályok száma alapján. A dolgozatom keretein belül felkutatok néhány létező megoldást, majd ezekre alapozva prezentálom saját megközelítésemet és ennek eredményeit. Az általam bemutatott megoldás egy multi-class multi-label osztályozást valósít meg mély neuronhálós rendszer segítségével többszólamú zenében.

1.1. Motiváció

Az ember kognitív képességei segítségével a zenében könnyedén fel tudja ismerni az egyes hangszereket. Ugyanez a feladat a számítógép számára azonban már sokkal kevésbé triviális. Ennek egyik oka, hogy egy hangszer megszólaltatásának digitális reprezentációja nagyon változatos lehet. Függ például a hangszíntől, hangmagasságtól, hangerőtől és előadásmódtól, de a felvétel minőségétől és az esetleges háttérzajtól is. További nehezítő körülmény a többszólamúság, amikor egy időben több hangszert is megszólaltatunk, ezzel összemosva az egyszólamú környezetben is sokváltozós képünket.

A MIR nagyban támaszkodik a mesterséges intelligenciára. A számítógépek számítási kapacitásának folyamatos növekedése és az elérhető adathalmazok gyarapodása által pedig egyre nagyobb figyelmet kap a mesterséges intelligencia egy kiemelten számításigényes részterülete: a mély tanulás. Ezt bizonyítja, hogy az évente megrendezésre kerülő ISMIR (International Society for Music Information Retrieval) konferencián 2010-ben még csak 2 ([1], [2]) mély tanulással kapcsolatos cikk jelent meg, de 2015-ben már 6, 2016-ban pedig már 16. [3]

A mély tanulás tehát egy ígéretes módszer lehet a MIR problémák megoldásában, ideértve az automatikus hangszerfelismerést is. Ezt kihasználva és megoldás életszerűségére törekedve döntöttem úgy, hogy elkezdek kísérletezni mély neuronhálókkal többszólamú zenében. Célom volt találni egy tanításra alkalmas adathalmazt, azon pedig tervezni egy olyan mély neuronhálós rendszert, amely a jelenlegi megoldások pontosságát meghaladja.

1.2. A dolgozat felépítése

Dolgozatomban tehát az eddigi kapcsolódó kutatásokat, illetve saját munkám eredményét dolgozom fel. A következő alfejezetben felsorolom az általam relevánsnak tartott, a State-of-the-Arthoz vezető kutatásokat.

A második fejezetben betekintést adok a téma elméleti hátterébe. Először kifejtem a zenével kapcsolatos főbb fogalmakat, bemutatom fontosabb tulajdonságait, reprezentációit. Kitérek a MIR bemutatására is. Ezután bevezetem a gépi tanulás és a mély tanulás fogalmát.

A harmadik fejezetben a módszertanról ejtek szót. Itt kifejtésre kerülnek az adatok előfeldolgozási módszerei, az általam bemutaott mély tanulási architektúrák, illetve ezek megvalósításai.

A negyedik fejezet az adathalmazokról fog szólni. Itt előbb felsorolom az adathalmazok kiválasztásának szempontjait, majd minden felhasznált adathalmaznak ismertetem a főbb jellemzőit.

Az ötödik fejezetben részletezem az általam végzett kisérleteket és ezek eredményeit. Ezeket összevetem egymással, illetve a releváns State-of-the-Art kutatásokkal.

A hatodik fejezetben összegzem a leírtakat, valamint továbbgondolom a kutatásomat, felvázolok néhány ötletet annak jövőjéről.

1.3. Kapcsolódó munkák

Az automatikus hangszerfelismerés témában a korábbi kutatások túlnyomó része a monofónikus, azaz egyhangszeres zenékkel foglalkozik. TODO!!!!!!

Többszólamú környezetbe való átültetéssel foglalkozott Burred tanulmánya [4], aki a többszólamúságot két kísérlettel közelítette meg. Először csak egy-egy hangjegyet kombinált össze többszólamú hangjeggyé. Itt két szimultán hangjegy esetén 73.15%-os, három hangjegyre 55.56%-os, négy hangjegy kombinációjára pedig 55.18%-os pontosságot sikerült elérni. Másik kísérletként hosszabb szekvenciákat kombináltak össze, ekkor két hang esetén 63.68%-os, három hang esetén pedig 56.40%-os pontosságot kaptak.

Salamon [5] egy hasonló témakörben tett kísérletet, polifónikus zenék dallamát kísérelték meg kinyerni. TODO!!!

Eggink és Brown [6] a polifónikus zenékben a hiányzó adat elméletükkel próbálták feltárni az egyes hangszereket. Ennek lényege, hogy felderítették azon idő- és frekvenciabeli részeket a zenén belül, ahol szeparáltan egy hangszer tulajdonságait vélték felfedezni és ezt dolgozták fel. Erre a módszerre épített Giannoulis és Klapuri [7] kutatása is, és hasonló megközelítést alkalmazott Garcia [8] is.

Jiang [9] egy többlépcsős megoldást mutat be. Első lépésben a hangszercsaládot határozták meg, ezzel szűkítve a a lehetséges hangszerek halmazát és a változók számát. A pontos hangszer-meghatározás csak ezután következett.

Az előbbi kutatások többnyire hagyományos gépi tanulási megoldásokat alkalmaztak, amelyekhez maguk nyerték ki a különböző bemeneti feature-öket. Humphrey [10] írásában a mély tanulási architektúrákat ismerteti a MIR terület korszerű irányzataként. A témában gyakorlati segítségként szolgál Choi [11] írása, amiben konkrét adatreprezentációkat, mély neuronhálós rétegeket, és mély tanulási technikákat mutat be.

Li [12] a nyers hanganyagot inputként felhasználva egy konvolúciós mély neuronhálós rendszert mutatott be a polifónikus zenében való automatikus hangszerfelismerés kapcsán. Ezt a megoldást aztán összevetette hagyományos gépi tanulási módszerekkel is. A mély neuronhálós rendszer teljesített legjobban. 75.60%-os pontossággal, 68.88%-os felidézéssel, 72.08 mikro F értékkel és 64.33 makro F értékkel. Han [13] szintén egy mély konvolúciós hálót használt, azonban az osztályozás szempontjából máshogyan járt el: a zenékben egy darab domináns hangszert keresett. Bemenetként a zenék spektogramját használta fel, 0.602-es mikro és 0.503-as makro F értéket ért el.

2. fejezet

Elméleti háttér

Ebben a fejezetben a dolgozathoz kapcsolódó fogalmakat és elméleti alapokat mutatom be. Először magának a zenének a releváns tulajdonságait mutatom be. Ezután ismertetem a MIR kutatási területet, amelybe dolgozatom is tartozik. Majd végül a mesterséges intelligencián alapuló megoldásokról nyújtok elméleti bevezetőt, érintve a hagyományos gépi tanulás és a mély tanulás módszereit is.

2.1. Zene és reprezentációi

2.1.1. Zene fogalma, tulajdonságai

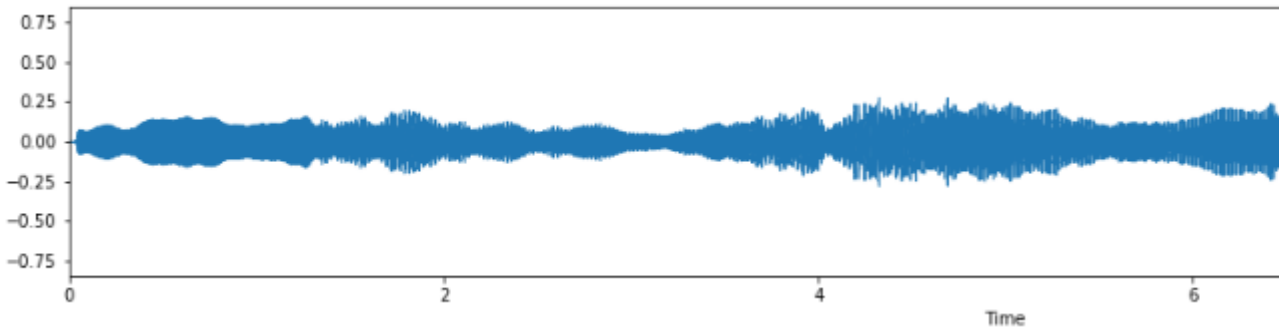
A zene egy meglehetősen összetett fogalom. Az ember számára a zene megjelenhet hang formájában, leírhatjuk őket szimbólumok segítségével egy kottában, előfordulhat szöveges formában dalszövegként, képi formában egy albumborító, vagy egy zenész képében, illetve mozdulatokban egy zenei előadás keretében.

Often music means the audio content, although otherwise its scope extends to other types of musical infor- mation e.g., lyrics, music metadata, or user listening history.

2.1.2. Hang reprezentációk

A hangok fizikai mivoltunkban rezgésekként jelennek meg. A rezgéseket matematikailag olyan folytonos függvényekkel tudjuk leírni, melyek értelmezési tartománya az idő, értékkészlete pedig a nyugalmi állapothoz viszonyított pillanatnyi kitérés.

Ilyen lehet például egy szinuszgörbe. Ahhoz, hogy a hangokat számítógépen tudjuk tárolni és feldolgozni, ezeket a függvényeket kell ábrázolnunk. Mivel azonban a számítógép számábrázolása véges, ezért a hangokat először digitalizálni kell. Ez azt jelenti, hogy a folytonos függvényeket diszkrét, azaz véges helyen vett és véges értékekkel rendelkező fügvényekké alakítjuk. Ez úgy történik, hogy ez eredeti függvényünkből megadott időközönként mintát veszünk, diszkrét értékre kerekítjük, és ezeket az értékeket összefűzzük. Az így kapott függvény lesz a hang digitális reprezentációja.



A hangok digitális reprezentációja tehát lényegében egydimenziós, mivel egy függvénygörbének tekinthetjük. Ezt szokták hívni nyers hangnak is, ugyanis további reprezentációkká tudjuk transzformálni. A MIR területen megjelenő mély tanulási megoldások jelentős része ezen nyers hangábrázolás helyett inkább a kétdimenziós reprezentációkat alkalmazza bemenetként. Ezt azzal indokolják, hogy a nyers bemeneten való tanítás sikeréhez nagyobb adathalmaz szükséges, mint a kétdimenziós reprezentációkéhoz. [11]

2.1.3. Music Information Retrieval (MIR)

A bevezetőben már említettem a zenei információk kinyerését (music information retrieval - MIR). Ez egy interdiszciplináris kutatási terület, magában hordozza többek között a zeneelmélet, pszichoakusztika, pszichológia, informatika, jelfeldolgozás és gépi tanulás tudományágakat. Céljára jól utal az elnevezése, zenékből szeretnénk releváns információt kinyerni, és ezeket felhasználni [11]. A felhasználásra szerintem nagyon jó, életszerű példát ad Downie 2003-as cikkének [14] bevezetője, amelyet a következőképp fordíthatunk le:

”Képzeljünk el egy világot, ahol egyszerűen felénekelhetjük egy számítógépnek a dalrészletet, ami már reggeli óta a fejünkben jár. A gép elfogadja a hamis énekün-

ket, kijavítja, és azonnal javaslatot tesz arra, hogy éppen a "Camptown Races" című számra gondoltunk. Miután mi belehallgatunk a gép által talált számos relevánsnak tartott MP3 fájl egyikébe, elfogadhatjuk a gép javaslatát. Ezután elégedetten elutasíthatjuk a felajánlást, hogy az összes további létező verzióját is felkutassa a dalnak, ide értve a nemrég megjelent olasz rap verziót, vagy a skótdudás duettre írt kottát." [14]

Figyeljük meg, hogy ez a hétköznapi eset mennyire össztett probléma. A következő feladatok jelennek meg:

- Az emberi éneklés, vagy dúdolás alapján hangfelismerés.
- Hang alapú lekérdezés egy zenei adatbázisban az előbbi bemenettel.
- Hangelemzés, feldolgozás, hogy a hamis hangokat ki tudjuk javítani, az esetleges háttérzajokat eltávolítsuk, illetve ha kell, a dallamból automatikusan kottát generáljunk.
- Hasonlóságon alapuló keresés zenék között, hogy megtaláljuk a kívánt dalt az adatbázisban.
- Zenei feldolgozások detektálása, hogy további verzióit is megtaláljuk egy adott dalnak.

MIR problémák definiálását több szempontból közelíthetjük meg. Choi cikkje [11] két tengelyre osztja fel a problémateret: szubjektivitás és eldöntési időmérték. A szubjektivitás tengelyen léteznek szubjektívebb feladatok, melyekre nincsenek egyértelmű válaszok. Ilyen lehet például a zene műfajának meghatározása. Objektívebb feladatoknak tekinthetjük azokat, melyek eredménye egyértelműen meghatárhozható, esetleg számszerűsíthető. Ide tartozik a hangszerfelismerés, vagy a tempó észlelés. [11]

A másik tengely, az eldöntési időmérték aszerint sorolja be a feladatokat, hogy mekkora időegységeken értelmezhető egy becslés. Ez egy relatív mérték. Például a dallamfelismerés eldöntési időmértékére azt mondhatjuk, hogy alacsony, mert egy felismert dallam jó eséllyel nem fedi le az egész zenét. Másik kifejezéssel azt mondhatjuk, hogy ez egy időben változó, azaz dinamikus tulajdonság. Ellenben a tempó

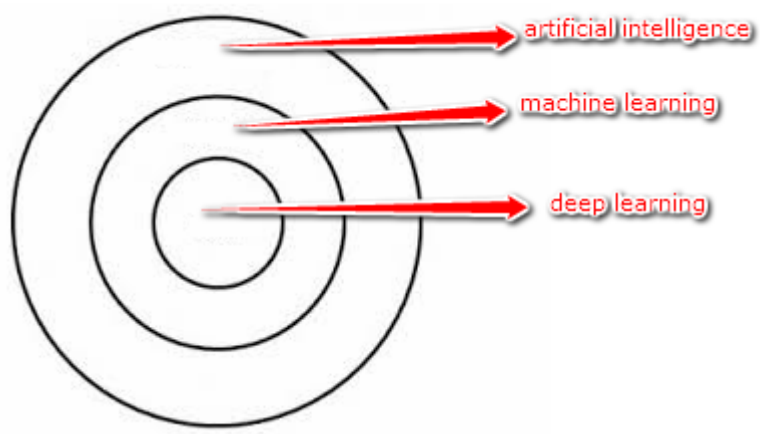
általában állandó értékű az egész zenében, így a teljes zeneszámot fel tudjuk címkézni egy adott tempóval. Erre azt mondjuk, hogy eldöntési időmértéke relatív magas, azaz ez egy statikus tulajdonsága a zenének.[11]

A hangszerfelismerést tekinthetjük statikus, illetve dinamikus feladatnak is a probléma megközelítésének függvényében. Dinamikus, ha erős címkézést szeretnénk megvalósítani, tehát arra vagyunk kíváncsiak, hogy adott időpillanatban éppen milyen hangszerek szólalnak meg. Gyenge címkézés esetén viszont a feladat statikussá válik. Ebben az esetben az egész zenére vetítve szeretnénk címkéket kapni egyes hangszerek jelenlétével, vagy a többi hangszerrel szembeni dominanciájával kapcsolatban.

TODO több MIR feladat

2.2. Machine learning, Deep Learning

ML, DL elhelyezése, 5. pdfem



2.2.1. Machine Learning

ML, ML in MIR

2.2.2. Deep Learning

DL, miben más mint az ML, architektúrák stb

3. fejezet

Módszertan

Lorem ipsum

3.1. Előfeldolgozás

adattisztítás, reprezentáció változtatás

3.2. Architektúra

3.3. Megvalósítás

3.3.1. Könyvtárak

4. fejezet

Adathalmaz

bevezető a datasetekről

4.1. Kiválasztási szempontok

pl. többszólam, ingyenesen elérhető, szerteágazó (not biased), stb

4.2. OpenMIC

4.3. MedleyDB

4.4. Egyebek

philharmonia orchestra irmas nsynth stb...

5. fejezet

Kísérletek, eredmények

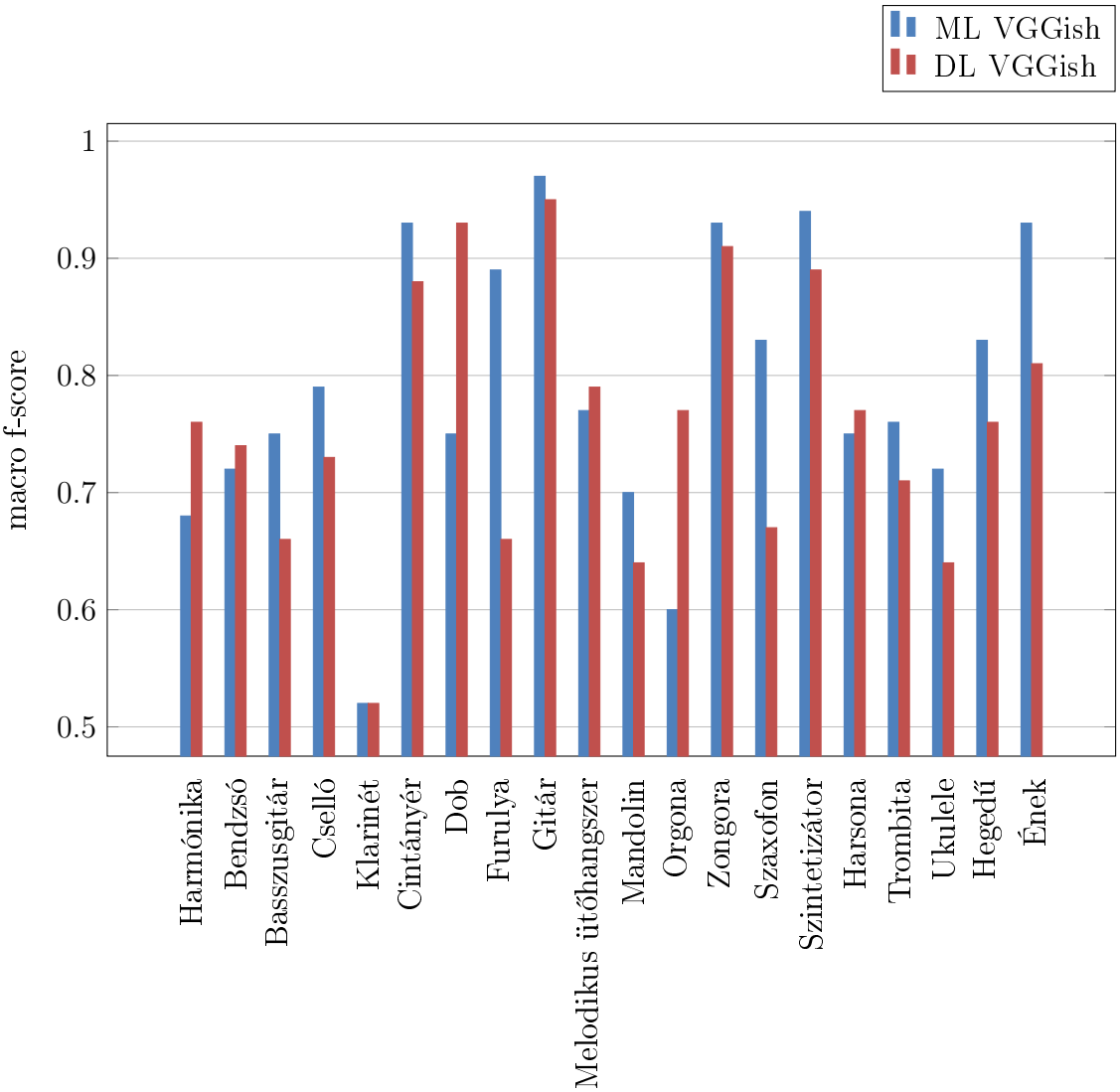
Lorem ipsum

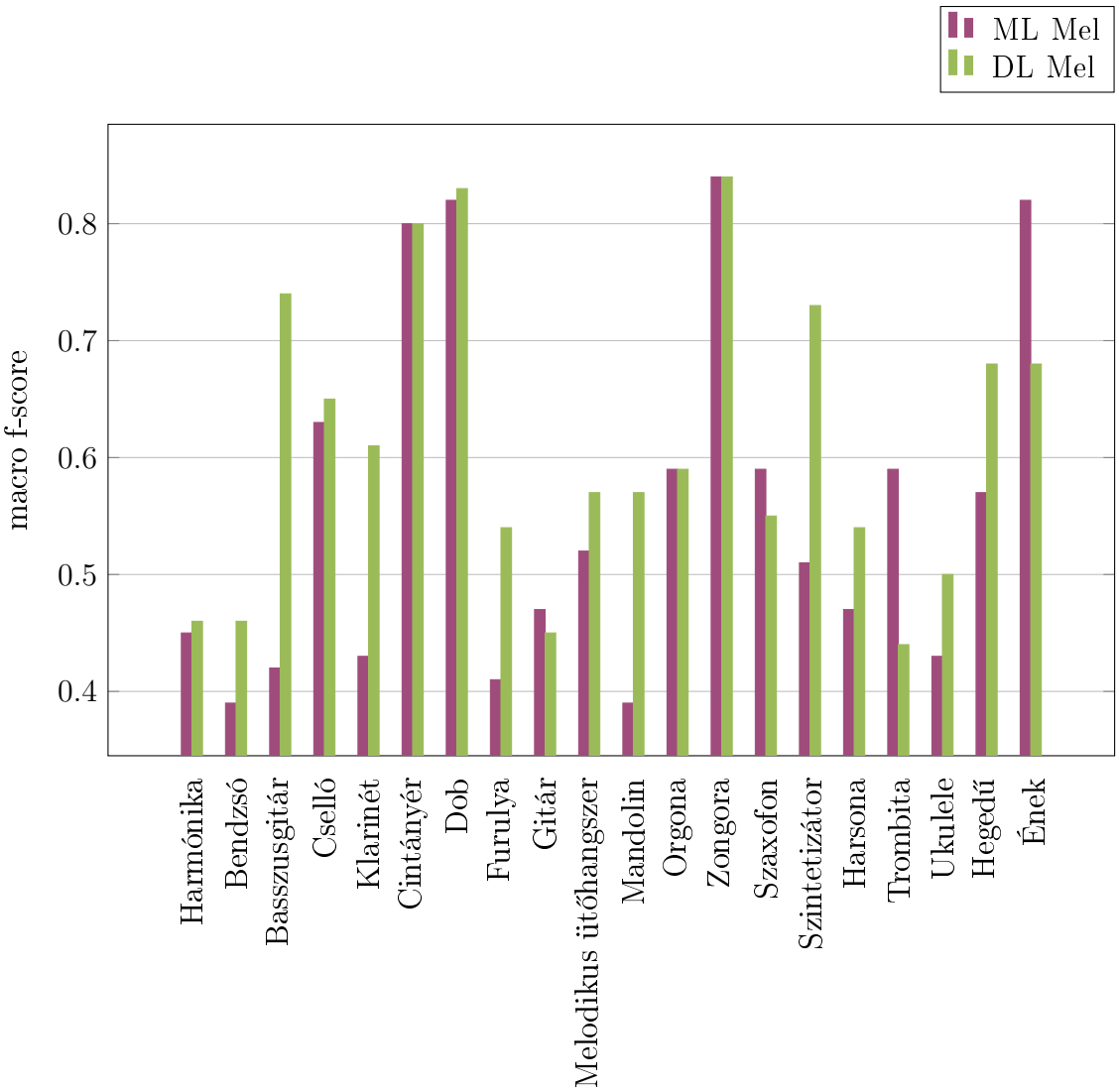
5.1. Mérőszámok

hogyan tudjuk megmérni a modell teljesítményét, pontosság, tanítás ideje, stb... A következő metrikákat vizsgáltuk:

- Pontosság (accuraccy) - a modell az adott lépésben a bemeneti adatok hány százalékára adott helyes kimenetet?
- Veszteség (loss) - a veszteségfüggvény eredménye. A modell predikcióinak a valóságtól való eltérését összeadva kapjuk meg.
- Precizitás (precision) -
- Felidézés (recall) -
- F1 érték (f1 score) - ..... Ennek a súlyozott átlaga a mérvadó.

5.2. Eredmények





5.2.1. próba1

5.2.2. próba2

5.2.3. próba3

5.2.4. próba4

5.3. Összehasonlítás

5.3.1. saját próbálkozásaim

5.3.2. más munkákkal (hagyományos ML)

6. fejezet

Összegzés, kitekintés

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In eu egestas mauris. Quisque nisl elit, varius in erat eu, dictum commodo lorem. Sed commodo libero et sem laoreet consectetur. Fusce ligula arcu, vestibulum et sodales vel, venenatis at velit. Aliquam erat volutpat. Proin condimentum accumsan velit id hendrerit. Cras egestas arcu quis felis placerat, ut sodales velit malesuada. Maecenas et turpis eu turpis placerat euismod. Maecenas a urna viverra, scelerisque nibh ut, malesuada ex.

Aliquam suscipit dignissim tempor. Praesent tortor libero, feugiat et tellus porttitor, malesuada eleifend felis. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nullam eleifend imperdiet lorem, sit amet imperdiet metus pellentesque vitae. Donec nec ligula urna. Aliquam bibendum tempor diam, sed lacinia eros dapibus id. Donec sed vehicula turpis. Aliquam hendrerit sed nulla vitae convallis. Etiam libero quam, pharetra ac est nec, sodales placerat augue. Praesent eu consequat purus.

Irodalomjegyzék

[1] Florian Eyben és tsai. "Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks." *Proceedings of the 11th International Society for Music Information Retrieval Conference* (Utrecht, Netherlands). Utrecht, Netherlands: ISMIR, 2010. aug., 589–594. old. DOI: 10.5281/zenodo.1417131. URL: https://doi.org/10.5281/zenodo.1417131.

[2] Philippe Hamel és Douglas Eck. "Learning features from music audio with deep belief networks". *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*. Utrecht, The Netherlands, 2010. aug., 339–344. old.

[3] Keunwoo Choi és tsai. "A tutorial on deep learning for music information retrieval". *arXiv preprint arXiv:1709.04396* (2017).

[4] Juan José Burred, Axel Roebel és Thomas Sikora. "Dynamic Spectral Envelope Modeling for Timbre Analysis of Musical Instrument Sounds". *Audio, Speech, and Language Processing, IEEE Transactions on* 18 (2010. ápr.), 663 –674. old. DOI: 10.1109/TASL.2009.2036300.

[5] J. Salamon és tsai. "Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges". *IEEE Signal Processing Magazine* 31.2 (2014), 118–134. old.

[6] J. Eggink és Guy Brown. "A missing feature approach to instrument identification in polyphonic music". 5. köt. 2003. nov., 49–. old. ISBN: 0-7803-7850-4. DOI: 10.1109/ICASSP.2003.1200029.

[7] Dimitrios Giannoulis és Anssi Klapuri. "Musical Instrument Recognition in Polyphonic Audio Using Missing Feature Approach". *Audio, Speech, and*

Language Processing, IEEE Transactions on 21 (2013. szept.), 1805–1817. old. DOI: 10.1109/TASL.2013.2248720.

[8] Jayme Barbedo és George Tzanetakis. "Musical Instrument Classification Using Individual Partials". *Audio, Speech, and Language Processing, IEEE Transactions on* 19 (2011. febr.), 111 –122. old. DOI: 10.1109/TASL.2010.2045186.

[9] Wenxin Jiang és Zbigniew Ras. "Multi-label automatic indexing of music by cascade classifiers". *Web Intelligence and Agent Systems* 11 (2013. ápr.), 149–170. old. DOI: 10.3233/WIA-130268.

[10] Eric J. Humphrey, Juan Pablo Bello és Yann LeCun. "Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics." *Proceedings of the 13th International Society for Music Information Retrieval Conference* (Porto, Portugal). Porto, Portugal: ISMIR, 2012. okt., 403–408. old. DOI: 10.5281/zenodo.1415726. URL: https://doi.org/10.5281/zenodo.1415726.

[11] Keunwoo Choi és tsai. "A Tutorial on Deep Learning for Music Information Retrieval". *CoRR* abs/1709.04396 (2017). arXiv: 1709.04396. URL: http://arxiv.org/abs/1709.04396.

[12] Peter Li, Jiyuan Qian és Tian Wang. "Automatic instrument recognition in polyphonic music using convolutional neural networks". *arXiv preprint arXiv:1511.05520* (2015).

[13] Yoonchang Han, Jaehun Kim és Kyogu Lee. "Deep convolutional neural networks for predominant instrument recognition in polyphonic music". *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (2016), 208–221. old.

[14] J. Stephen Downie. "Music information retrieval". *Annual Review of Information Science and Technology* 37.1 (2003), 295–340. old. DOI: 10.1002/aris.1440370108. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440370108. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370108.