GENERATÍV HANGMODELLEZÉS

Szökrön Dorottya - Önálló projekt III. (2022/23) - Projektbeszámoló

2022. december 19.

Tartalomjegyzék

1. Bevezetés

Ebben a félévben új típusú neurális hálózatokkal, GAN-okkal kezdtünk el foglalkozni, és a félév elején ennek az elméleti hátterével ismerkedtem meg. Úgy gondoltuk, hogy a korábbi fehérjeklasszifikációs feladat helyett a hangmodellezés területén használjuk fel ezeket a gyakorlati munka során. A témában megismert modellek emberi beszédszintézisre lettek kifejlesztve, viszont mi zenei hangszintézisre használtuk fel őket. A generált hangminták megfelelő értékelése miatt bővebben foglalkoztunk mel-spectogramokkal, és az azonos kiindulási hangmintákhoz tartozó különböző generátumok közötti lehetséges metrikával. A feladathoz használt adathalmazok kialakítására is fordítottunk időt, melyek CD minőségű, több órás hangszeres felvételek feldolgozásával jöttek létre.

A GAN-ok alkalmazása egy olyan alkotói folyamatnak képzelhető el, mint például egy portré rajzolása. Emiatt nehezebbnek mondható a mély gépi tanulási feladatok között, hiszen könnyebben deklarálható egy klasszifikációs feladat és a hozzá tartozó feltételhalmaz, mint egy kreatív alkotói folyamat meghatározása. Viszont az ilyen típusú feladatok megoldása közelebb visz az emberi intelligencia megértéséhez, illetve reprodukálásához.

2. Kapcsolódó fogalmak

Hang alaptulajdonságai

Fizikai értelemben a különböző hangjeleket a légnyomás változása hozza létre. A nyomásváltozások intenzitása mérhető, és ezek a mérések az idő függvényében ábrázolhatóak. Ha ezek a jelek rendszeres időközönként ismétlődnek, akkor minden hanghullám azonos alakú lesz. A hullámok magassága a hang intenzitását mutatja, és amplitúdónak nevezik. A periódus az az idő, amely alatt a jel egy teljes hullámot teljesít, a frekvencia pedig a jel által egy másodperc alatt keltett hullámok számát jelenti, és mértékegysége a Hertz. Tehát a frekvencia a periódus reciproka. A legtöbb hang, amellyel találkozunk, nem ilyen egyszerű és szabályos periodikus mintákat követ. De a különböző frekvenciájú jelek összeadhatók és így összetettebb, ismétlődő mintázatú jelek jönnek létre. Minden hang ilyen hullámokból áll, és az emberi fül képes megkülönböztetni a különböző hangokat a benne szereplő frekvenciák összessége és változásuk alapján, amely a hangszínnek definiálható.

A hangok digitális ábrázolása

Egy hanghullám digitalizálásához a jelek számsorokká alakítása szükséges, ez a hang amplitúdójának meghatározott időközönkénti mérésével történik. Minden ilyen mérés egyetlen minta lesz, és a mintavételi sebesség a másodpercenkénti minták számát jelöli. Például a hifi minpségű mintavételi sebesség 44100 minta másodpercenként.

Azonban a mély tanulási algoritmusok során jellemzően nem a hangadatok, nyers formája kerül felhasználásra, hanem a hangjelek képekké alakítása történik. Ez a hangból spektrogramok generálásával valósul meg. Majd hagyományos konvolúciós hálóarchitektúrák felhasználásával kerülnek a képek feldolgozásra.

Spektrum

A különböző frekvenciájú jelek összeadásával a való életben előforduló bármely hang reprezentálható. Ez azt jelenti, hogy egy jel több különböző frekvenciából állítható össze, és ezeknek a frekvenciáknak az összegeként fejezhető ki. A spektrum azon frekvenciák összessége, amelyek kombinálásával a jel elő lett állítva. Tehát a spektrum ábrázolja a jelben lévő összes frekvenciát az egyes frekvenciák amplitúdójával együtt.

A hangjel ábrázolásának egyik módja az amplitúdót az idő függvényében mutatni. Ebben az esetben az x tengely a jel időértékeinek tartománya, tehát a jel az időtartományban van tekintve. A spektrum egy alternatív módja a jelek az ábrázolásának, mert az amplitúdót a frekvencia függvényében mutatja, és ekkor az x tengely a jel frekvenciaértékeinek tartománya, egy adott pillanatban, tehát a frekvenciatartományban van tekintve a jel.

Spektrogram

Egy jel különböző hangokból tevődik össze, ezért az idő múlásával változik, és így az alkotó frekvenciái is az idővel változóak. Tehát a spektruma is változik az idővel. Egy jel spektrogramja az idő függvényében ábrázolja a spektrumát, és úgy képzelhető el, mint a jel „fényképe”. Az x tengelyen az időt, az y tengelyen pedig a frekvenciát jeleníti meg. Olyan, mintha különböző időpontokban megtörténne a spektrum ábrázolása majd ezek az ábrák együttesen lennének vizualizálva. A spektogramon különböző színek felelnek meg az egyes frekvenciák amplitúdójának. Minél világosabb a szín, annál nagyobb a jel amplitúdója. A spektrogram minden egyes függőleges „szelete” lényegében a jel spektruma az adott pillanatban, és

megmutatja, hogy a jelerősség hogyan oszlik el a jelben az abban a pillanatban megtalálható minden frekvencián.

Spektrogramok a hangjelekből Fourier-transzformáció alkalmazásával állíthatók elő, ezzel a módszerrel lehetséges bármilyen jelet az azt alkotó frekvenciákra bontani és megjeleníteni a jelben jelenlévő egyes frekvenciák amplitúdóját. A hangjel időtartamának kisebb időszegmensekre felvágása után minden szegmensre alkalmazni kell a Fourier-transzformációt, hogy meg legyenek határozva az adott szegmensben lévő frekvenciák. Ezután az összes szegmens Fourier-transzformációit egyetlen diagram egyesíti.

Mel-skála

Elvégezve az eljárást és ábrázolva a kapott spektogramot, megfigyelhető, hogy nem olvasható le róla sok információ emberi szemmel. Ez azzal a jelenséggel magyarázható, hogy az emberi hallás a frekvenciák és amplitúdók csak egy szűk tartományára összpontosul. Hangmagasságnak nevezik azt, ahogyan az emberek hallják a frekvenciákat, ez egy szubjektív impresszió. Bár a magas hangnak magasabb a frekvenciája, mint az alacsony hangnak, azonban az emberek nem lineárisan érzékelik a frekvenciákat, hanem érzékenyebbek az alacsonyabb frekvenciák közötti különbségekre, mint a magasabb frekvenciák közöttiekre. Például 100 Hz-es és 200 Hz-es hangminták között nagyobb különbség hallható, mint az 1000 Hz-es és 1100 Hz-es minták közöt, és kevesen tudnak különbséget tenni 10000 Hz-es és 10100 Hz-es minták között. Tehát megállapítható, hogy az emberek inkább logaritmikus skálán hallják a hangokat, mint lineáris skálán.

A Mel-skálát azért fejlesztették ki, hogy ezt a jelenséget is figyelembe lehessen venni. Ehhez nagyszámú résztvevővel végeztek kísérleteket, úgy hogy mintákat hallgattattak velük. Így hozták létre ezt a hangmagasság-skálát, olyan beosztás szerint, hogy a hallgatók minden egységet egyenlő hangmagasság-távolságnak ítéltek meg a rákövetkezőtől.

Decibel skála

A hang amplitúdójának emberi érzékelése a hang erőssége. Hasonlóan a frekvenciához, a hangerősségét inkább logaritmikusan, mint lineárisan halljuk. Ezért fejlesztették ki a Decibel-skálát a Mel-skálához hasonlóan. Ezen a skálán 0 dB a hallásküszöbi hangzás majd a mértékek exponenciálisan nőnek. Tehát a 10 dB tízszer hangosabb, mint a 0 dB, a 20 dB 100-szor, a 30 dB pedig az 1000-szer. A 120 dB a fájdalom küszöb, e feletti hang már az emberi fülnek elviselhetetlenül hangosnak mondható.

Mindezek alapján látható, hogy a hang valósághű ábrázolásához a Mel-skála és a Decibel-skála használata szükséges, amikor az adatainkban szereplő frekvenciákkal és amplitúdókkal foglalkozunk.

Mel-spektrogram

Az előzőek alapján figyelembe vett szempontok szerinti ábrázolási eszköz a Mel-spektrogram. Két fontos különbség van eközött és egy szokásos spektrogram között, amely a frekvencia és az idő függvényét ábrázolja. Első, hogy az y tengelyen a frekvenciát a Mel-skála, a második, hogy a színek megadásában az amplitúdót a Decibel-skála méri.

A mély tanulási modelleknél általában Mel-spektogram alkalmazása történik az egyszerű spektrogram helyett. A Mel-spektrogramok előállításához alkalmazott Fourier-transzformációk kiszámításának egyik módja a diszkrét Fourier-transzformáció nevű módszer, a gyakorlatban a gyors Fourier Transzformáció (FFT) algoritmust használják. Az FFT megadja az általános frekvenciakomponenseket az audiojel teljes időtartamára vonatkozóan. Tehát az nem derül ki, hogyan változnak ezek a frekvenciakomponensek a jelen belül az idő múlásával.

A részletesebb infromációk, például a frekvencia időbeli változásainak kinyerésére a Short-Time Fourier Transzformáció (STFT) algoritmus alkalmas. Az STFT a Fourier-transzformáció egy olyan változata, amely a hangjelet meghatározott méretű (általában nem diszjunkt) időintervallumokra bontja. Minden szakaszon kiszámítja az FFT értékét, majd egyesíti ezeket, így képes információt adni a frekvencia időbeli változásairól.

Tehát a jel először szakaszokra osztódik az x (idő) tengely mentén, majd az y (frekvencia) tengely mentén is. Ezután minden egyes időszakaszra kiszámítódik az amplitúdó a frekvenciasávokhoz, amelyek a Mel-skálát figyelembevéve fognak alakulni.

Legyen egy 1 perces hangfelvétel, amely 0 és 10000 Hz közötti frekvenciákat tartalmaz (Mel-skálán). A Mel-spektrogram algoritmus intervallum méret paramétere legyen akkora, hogy a hangjel 20 időszakaszra legyen bontva. Valamint a frekvenciatartomány legyen 10 sávra osztva (azaz $0 - 1000$ Hz, $1000 - 2000$ Hz, $\ldots$, $9000 - 10000$ Hz).

Az algoritmus kimenete egy 2-dimenziós tömb lesz (10, 20), amelyben a 20 oszlop felel meg egy-egy időszakaszra alkalmazott FFT algoritmus kimenetének, és a 10 sor egy-egy frekvenciasáv amplitúdóértékét reprezentálja. Például az első oszlop, ami az első időintervallumra alkalmazott FFT 10 sora közül az első sor az első frekvenciasáv amplitúdója $0 - 1000$ Hz között, a második sor a második frekvenciasáv amplitúdója $1000 - 2000$ Hz között és hasonlóan folytatódik. Az így kapott tömb minden oszlopa megfelel egy sávnak a mel spektrogram képén.

Mel-spektrogram hiperparaméterei

Az előzőek alapján a mel spektrogramok előállításához különböző paraméterek meghatározása szükséges. A Python Librosa könyvtára által használt paraméterneveket fogom használni.

Frekvenciasávok

- f_{min} - minimális frekvencia
- f_{max} - a megjelenítendő maximális frekvencia
- n_{mels} - a frekvenciasávok száma (a spektrogram magassága)

Időszakaszok

- n_{fft} - időintervallumok hossza
- hop_{length} - azon minták száma, amennyivel az intervallumot minden lépésnél el kell csúsztatni, tehát a spektrogram szélessége = minták száma összesen / hop_{length}

Generatív modellezés

Felügyelt tanulás során a cél az x bemeneti adatok és az azokhoz tartozó y kimeneti adatok közötti leképezés megtanulása, a bemenet-kimenet párok címkézett halmaza alapján. Ilyen típusú problémák például az osztályozási probléma és a regresszió feladat, a felügyelt tanulási algoritmusok közé pedig a logisztikus regresszió és a random-forest tartozik. Ezzel szemben a felügyelet nélküli tanulás során a modell csak az x bemeneti változókat kapja meg, és ezekhez nem tartoznak y kimeneti változók. A cél érdekes mintázatok felfedezése az input adatokban. A nem felügyelt tanulási problémák például a klaszterezés és a generatív modellezés, és az ehhez tartozó algoritmusok a k-közép módszer és a generatív ellenséges hálózatok.

Generatív modellezésnek az a felügyelet nélküli tanulási módszer (*unsupervised learning*) nevezhető a gépi tanulás területén, amely magába foglalja a bemeneti adatok szabályszerűségeinek, mintázatainak automatikus felfedezését és megtanulását oly módon, hogy a modell később felhasználható legyen olyan új példák generálására, amelyek az eredeti adatkészletből származónak tűnhetnek.

A generatív ellenséges hálózatok (*Generative Adversarial Networks (GAN)*) a generatív modellezés egyik megközelítése mély tanulási módszerekkel, például konvolúciós neurális hálózatokkal. A GAN-ok alkalmazásával a probléma felügyelt tanulási feladatként fogalmazható meg két alapmodell segítségével:

- generátor modell: új példák generálását tanulja meg, amelyek hasonlóak az eredeti adathalmaz elemeihez,
- diszkriminátor modell: a példákat valósnak (*from the domain*) vagy hamisnak (*generated*) klasszifikálja.

A két modell tanítása együttesen történik egy zéró összegű, úgymond ellenséges játékban, ami azt jelenti, hogy a játékosok a saját nyereségüket csak a másik játékos kárára növelhetik. Ebben az esetben a zéró összeg azt jelenti, hogy ha a diszkriminátor sikeresen osztályozza a valódi és hamis példákat, akkor jutalmat kap, nem lesz szükséges a modell paramétereinek módosítása, míg ezzel szemben a generátor büntetése a modellparaméterek frissítése lesz. Vagy hasonlóan, ha a generátor meg tudja téveszteni a diszkriminátort, jutalmat kap, tehát nincs szükség a modell paramétereinek módosítására, de a diszkriminátor büntetése, hogy frissíteni kell a modell paramétereit.

Egy határ után a generátor minden alkalommal olyan generált példákat fog létrehozni, hogy a diszkriminátor nem fogja tudni megkülönböztetni az eredeti adatoktól, és minden esetben bizonytalanul, azaz véletlenszerűen fog jósolni ($50 - 50\%$, hogy valódi vagy hamis). A folyamat ideális esetben eddig tart, azonban nem feltétlenül kell idáig eljutnunk ahhoz, hogy egy hasznos generátormodellhez jussunk.

Diszkriminátor modell

A diszkriminátor egy osztályozási modell lesz, amelynek architektúrája bármilyen lehet, ami megfelel az általa osztályozott adattípusnak. Ezt a modellt a tanítási folyamat után sok esetben nem tartjuk meg. A diszkriminátor tanító adathalmaza kétféle forrásból származik. Egy része a valódi adathalmaz elemeiből áll elő, amelyek a valós osztályba vannak sorolva a tanítás során. Valamint a generátor által létrehozott hamis példákból, amelyek a hamis osztályba vannak sorolva a tanítás során.

A diszkriminátor tanítási folyamata alatt a generátor passzív állapotban van, azaz a modell nem tanul, tehát a súlyai állandóak maradnak, miközben a példákat készíti el a diszkriminátor számára.

A diszkriminátor két veszteségfüggvényhez kapcsolódik, de a tanítási folyamat során figyelmen kívül hagyja a generátor veszteségfüggvényét, és csak a sajátját használja fel.

Tanítási folyamat:

- valós és hamis adatok osztályozása,
- a veszteségfügvény megbünteti a diszkriminátort, ha egy valós mintát hamisnak, vagy ha egy hamis mintát valósnak minősít
- a diszkriminátor modell súlyainak frissitése backpropagation algoritmussal

A következő részben kerül kifejtésre, hogy a generátor veszteségfüggvénye miért kapcsolódik a diszkriminátorhoz.

Generátor modell

A generátor modell megtanul hamis adatokat létrehozni a diszkriminátortól származó visszajelzéseket is felhasználva, tehát olyan példákat generálni, amelyeket a diszkriminátor valósnak minősít.

A generátor modell egy rögzített hosszúságú véletlen vektort kap bemenetként, és ebből generálja a mintát. Alapvetően ez egy véletlenszerű zaj, amit a generátor értelmes kimenetté alakít. A véletlen zaj alkalmazásával azt érhetjük el, hogy a modell sokféle adatot állítson elő. Tapasztalatok alapján a zaj elolszlása nem befolyásoló tényező, ezért általában normális-eloszlásból vesszük.

A generátor modell tanítása során szoros kapcsolat van a generátor és a diszkriminátor között. Ahhoz, hogy egy háló veszteségét csökkenteni tudjuk a súlyokat kell megváltoztatni, a generátor veszteségfüggvénye megbünteti a generátort, ha olyan példát állít elő, amit a diszkriminátor hamisnak minősít. Viszont a hálózatban a generátor nem kapcsolódik közvetlenül a veszteségfüggvényhez, csak a diszkriminátoron keresztül. Ez utóbbi súlyait nem változtathatjuk meg a tanítás során, ezért a generátor tanítási folyamata a következőképpen történik.

- véletlenszerű zajminta vétele
- a generátor kimenetet állít elő a mintavételezett zajból
- a diszkriminátor osztályozásának eredménye az előállított kimenetre
- veszteség kiszámítása az eredmény alapján
- backpropagation algoritmus alkalmazása a diszkriminátoron, és a generátoron keresztül is
- a kapott gradiens értékekkel csak a generátor súlyainak módosítása

Tekintve az egész GAN hálózat tanítási folyamatát, ekkor két fázis különböztethető meg. Az első fázisban a diszkriminátor tanítása folyik egy vagy több epochon keresztül, a második fázisban pedig a generátoré. Ezen fázisok egymás utáni váltakozása a GAN tanítása.

3. HiFi-GAN

Általánosan elterjedt módszer a beszédszintézis feladat megoldására autoregressive (AR) és flow-based generatív modellek használata, de a közelmúlbeli kutatások során elkezdtek GAN hálózatokat is alkalmazni. Ez a módszer javította a mintavételi hatékonyságot és a memóriahasználatot, azonban a generált példák minőségének szempontjából rosszabb eredményt értek el vele, mint a korábbi modellekkel. Az általam feldolgozott cikkben bemutatott HiFi-GAN modell a szerzők kísérletei alapján nagyobb számítási hatékonyságot és mintaminőséget képes elérni, mint az AR vagy flow-based modellek.

A HiFi-GAN modell egy generátorból és két diszkriminátorból, egy multi-scale diszkriminátorból (MSD) és egy multi-period diszkriminátorból (MPD) áll.

Generátor

A generátor egy teljesen konvolúciós (*fully convolutional*) neurális hálózat, ami olyan konvolúciós hálózatot jelent, amely nem tartalmaz teljesen kapcsolt (*fully connected*) réteget. Mel spektrogramot kap bemenetként, és transzponált konvolúciókon keresztül dolgozza fel azt. Minden transzponált konvolúciót egy multi-receptive field fusion (MRF) modul követ.

MRF modul

Párhuzamosan dolgozza fel a különböző hosszúságú mintákat, úgy hogy a modul reziduális blokkok kimeneteinek összegét adja vissza. Különböző kernel size és dilation rate paraméterek vannak meghatározva mindegyik blokkhoz, így különböző receptív mezők (*receptive field*) alakulnak ki. A következő paraméterek határozhatóak meg

- h_u - transzponált konvolúció rejtett dimenziója
- k_u - transzponált konvolúció során alkalmazott kernelméret
- k_r - MRF modulban alkalmazott kernelméret
- D_r - MRF modulban alkalmazott dilatációs érték

MPD

Az MPD több aldiszkriminátorból áll, amelyek mindegyike a bemeneti hang periodikus jeleinek csak egy részét dolgozza fel, úgy hogy az eredeti hangjel minden p. mintáját tekinti csak. Tehát először a T hosszúságú 1-dimenziós nyers hangot T/p magasságú és p szélességű 2-dimenziós adatokká alakítja át, majd 2-dimenziós konvolúciókat alkalmaz. Az MPD minden konvolúciós rétegében a kernel mérete 1 azért, hogy a periodikus minták függetlenül legyenek feldolgozva, és ReLU aktivációs függvényt használ.

MSD

Az MSD három aldiszkriminátorból áll, amelyek különböző bemeneti adatokat kapnak, az egyik a nyers hangot, a másik a kétszeresen és átlagosan összevont hangot, a harmadik a négyszeresen és átlagosan összevont hangot. Mivel az MPD minden aldiszkriminátora csak diszjunkt mintákat fogad el, összeadhatjuk az MSD-vel az audioszekvencia egymás utáni kiértékeléséhez. Az MSD minden aldiszkriminátora szintén ReLU aktivációs függvényt használ.

Veszteségfüggvény

Egy GAN veszteségfügvényei a G generátort és a D diszkriminátort tekintve a következőképpen definiálhatóak:

$$\mathcal{L}_{Adv}(D;G) = \mathbf{E}_{(x,s)}\Big[(D(x)-1)^2 + (D(G(s)))^2\Big]$$

$$\mathcal{L}_{Adv}(G;D) = \mathbf{E}_{s}\Big[(D(G(s))-1)^2\Big]$$

ahol x az eredeti hangjel, s ennek a mel spektogramja.

Ezeken kívűl még definiálni szükséges a mel spektrogram veszteséget:

$$\mathcal{L}_{Mel}(G) = \mathbf{E}_{(x,s)}\Big[||\varphi(x) - \varphi(G(s))||_1\Big]$$

ahol φ a mel spektrogramot előállító függvény, és L_1 norma van alkalmazva.

A generátor tanításához az alábbi veszteségfüggvényt kell alkalmazni:

$$\mathcal{L}_{FM}(G;D) = \mathbf{E}_{(x,s)}\Big[\sum_{i=1}^{T}\frac{1}{N_i}||D^i(x) - D^i(G(s))||_1\Big]$$

ahol T jelöli a rétegek számát a diszkriminátorban, D^i és N_i jelölik a diszkriminátor i-edik rétegének jellemzőit, illetve a jellemzők számát.

Tehát végül a modell veszteségfügvényei a G generátort és a D diszkriminátort tekintve a következőképpen állíthatóak elő:

$$\mathcal{L}_G = \mathcal{L}_{Adv}(G;D) + \lambda_{fm}\mathcal{L}_{FM}(G;D) + \lambda_{mel}\mathcal{L}_{Mel}(G)$$

$$\mathcal{L}_D = \mathcal{L}_{Adv}(D;G)$$

a $\lambda_{fm} = 2$ és a $\lambda_{mel} = 45$ értékek vannak megadva a HiFi-GAN modell esetében.

Végül definiálhatóak a veszteségfüggvények az aldiszkriminátorokra is:

$$\mathcal{L}_G = \sum_{k=1}^{K}\Big[\mathcal{L}_{Adv}(G;D_k) + \lambda_{fm}\mathcal{L}_{FM}(G;D_k)\Big] + \lambda_{mel}\mathcal{L}_{Mel}(G)$$

$$\mathcal{L}_D = \sum_{k=1}^{K}\mathcal{L}_{Adv}(D_k;G)$$

ahol D_k jelöli a k-adik aldiszkriminátort az MPD-ben valamint az MSD-ben.

4. Adathalmaz

Beszédszíntézis

A HiFi-GAN modell eredetileg beszédszintézis feladat megoldására lett kifejlesztve és optimalizálva. Az ehhez felhasznált adathalmaz az LJ Speech Dataset. Ez egy nyilvánosan elérhető beszédadatkészlet, amely 13100 rövid hangfelvételt tartalmaz. Ezek mindegyike egyetlen beszélőtől származik, aki 7 nem szépirodalmi könyvből olvas fel részeket. A felvételek hossza 1 és 10 másodperc között változik, az összes hangadat teljes hossza pedig körülbelül 24 óra.

Hangszintézis

A projektben a célunk a HiFi-GAN beszédszintézisre fejlesztett modell felhasználása zenei hangszintézisre, tehát klasszikus hangszerek hangjának generálása. Az ehhez szükséges klasszikus zenei adathalmaz kialakítását magam végeztem el. Egy publikus zenemegosztó oldalról választottam hangfelvételeket majd alakítottam át a modellnek megfelelő formátumra. A hangadatok kiválasztása több szempont szerint történt, az egyik szempont a hangszer fajtája, másik szempont a hangszerek mennyisége volt. Ezek alapján 5 db körülbelül másfél órás hangfelvételt használtunk fel a munka során.

- Piano - az adathalmaz kizárólag zongora darabokat tartalmaz ismeretlen szerzőktől
- Bach - a zeneszerző összes fuvolaszonátáját tartalmazza, Emmanuel Pahud fuvolaművész előadásában, Trevor Pinnock csembaló kíséretével
- Vivaldi - a zeneszerző egyik hegedűversenyének összes tétele két hegedűművész (Piero Toso, Juan Carlos Rybin) és egy oboaművész (Pierre Pierlot) előadásában
- Tchaikovsky - Diótörő című balettjének zenei felvétele leszűkítve körülbelül másfél órára, a mű hangszerelése: I. hegedű, II. hegedű, brácsa, cselló, nagybőgő, 2 fuvola, 2 oboa, piccolo, angolkürt, 2 klarinét, 2 fagott, 4 kürt, 2 kornett, 2 trombita, 3 harsona, tuba, triangulum, tamburin, timpani, cimbalom, pergődob, dob, harangjáték, tam-tam, cseleszta, kasztanyetta, játékhangok, 2 hárfa, zongora
- Waltzes - Strauss, Nino Rota, Tchaikovsky, Chopin zeneszerzők válogatott keringő darabjai, különféle hangszerek megszólaltatásában, előfordul köztük csak zongorán és nagyobb zenekarok által előadott felvétel is

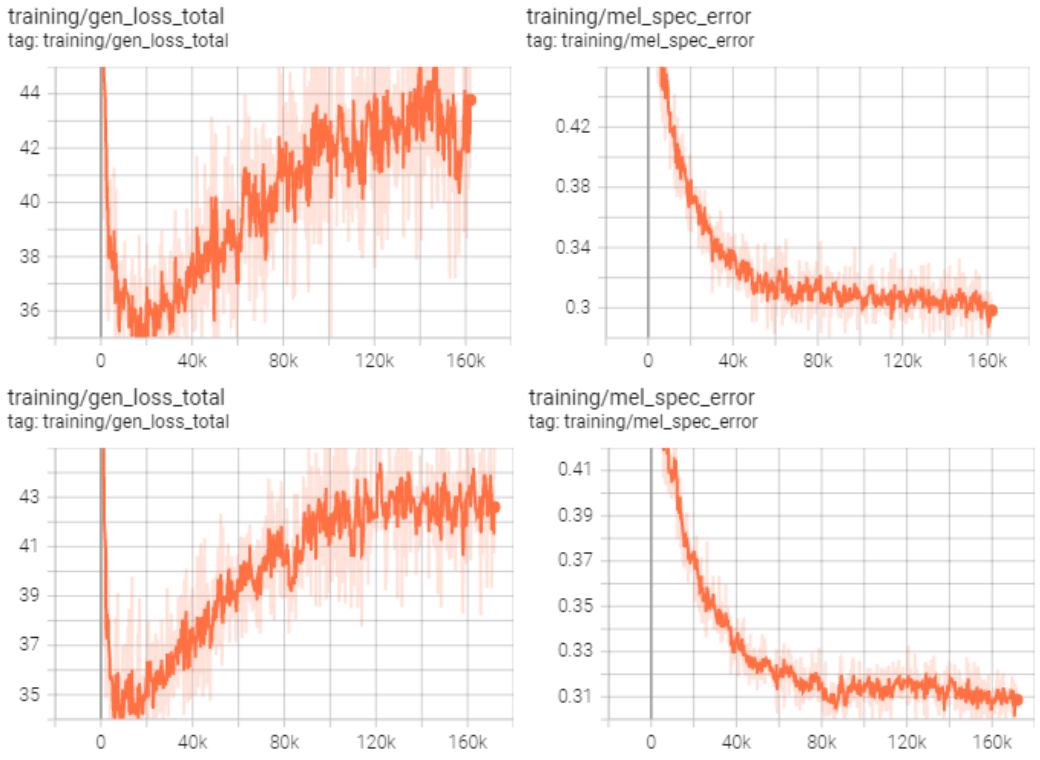
Mindegyik felvételt *.wav* formátumúvá alakítottam át és 10 mp-es részletekre osztottam fel.

5. Eredmények

A feladat sajátosságaiból adódóan az eredmények reprezentálása bővebb kereteket igényel, mint egy írásos formátum, ezért létrehoztam egy kísérő weboldalt, ahol a különböző tanítási fázisok után generált hangmintákat lehet meghallgatni.

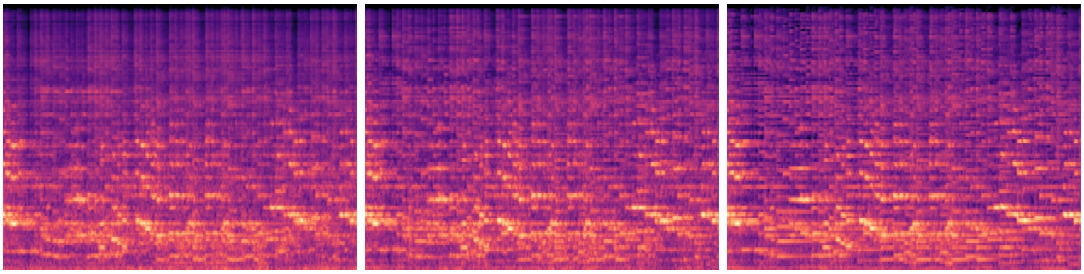
`https://szokrondorka.github.io/`

Viszont nem elegendő csak emberi érzékelés segítségével értékelni a modell teljesítményét, ezért elvégeztük a különböző generátumokhoz tartozó mel spektogramok és az eredeti hangadathoz tartozó mel spektogramok összehasonlítását. Az spektrogramok RGB színtérben vannak elkódolva. Ez egy háromdimenziós tér, melynek koordinátái rendre a piros, zöld, kék alapszínek intenzitását határozzák meg.



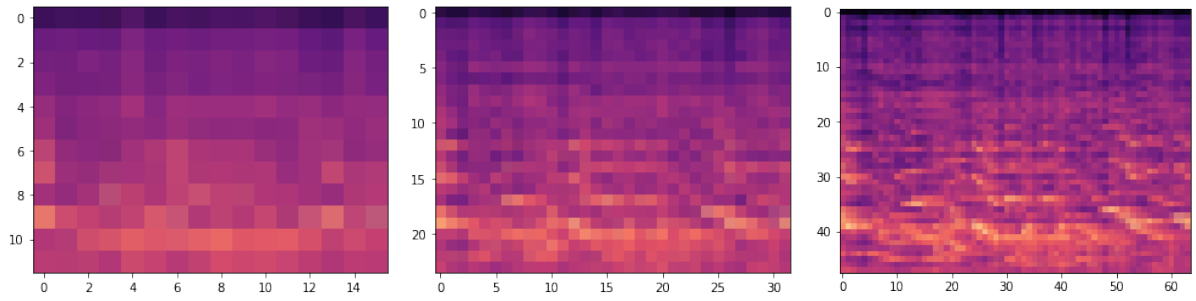
1. ábra. A modell tanulási görbéje a Bach (felül) és a Tchaikovsky (alul) adathalmazon 5000 epochon keresztül

Az alábbi ábrán megfigyelhető a mel spektogram képének változása a tanítás során. Bár szabad szemmel is láthatóak a különbségek célszerűbb valamilyen távolságot bevezetni a spektogramok között.



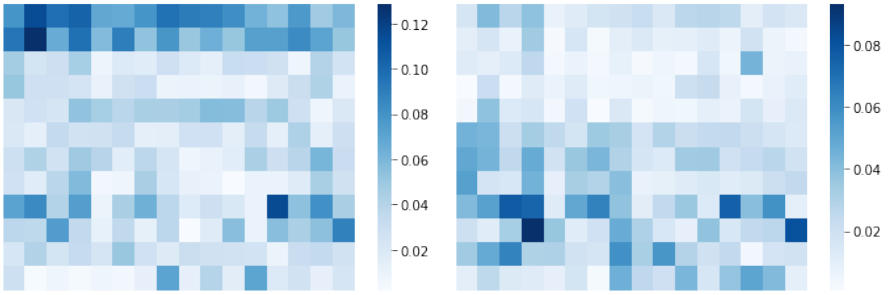
2. ábra. Jobbról balra haladva 500 epoch után, 5000 eopch után, az eredeti hang mel spektogramja

A fent definiált színtérben többféleképpen tudunk távolságot mérni, én a kísérletek során L_1 és L_2 normát használtam. Először aggregáltam a képeket 40, 20 és 10 pixelenként, tehát $n * n$-es négyzetekre osztottam fel a képeket és egy 2-dimenziós average pooling réteg segítségével átlagoltam az értékeket. Később ezeket az aggregált ábrákat használtam fel az összehasonlításokhoz. Az alábbi ábrán az eredeti hangjel aggregált mel spektrogramjai láthatóak.

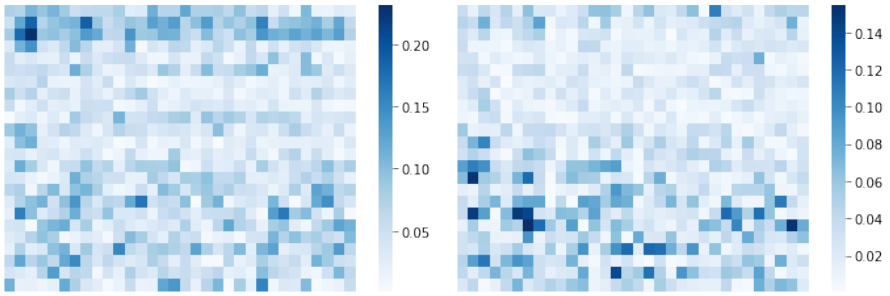


3. ábra. Jobbról balra haladva 40, 20 és 10 pixelenként aggregálva

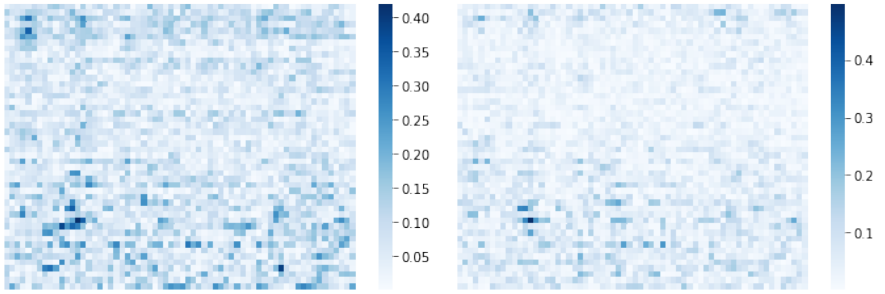
A különbségek megjelenítéséhez a hőtérképes módszert választottam, melyen minél sötétebb a szín, annál nagyobb távolságot jelent. A következő ábrákon a különböző felosztások után és a két normában kapott eredmények figyelhetőek meg. Az elvárásoknak megfelelően a felosztás finomításával pontosabb információkat kaphatunk.



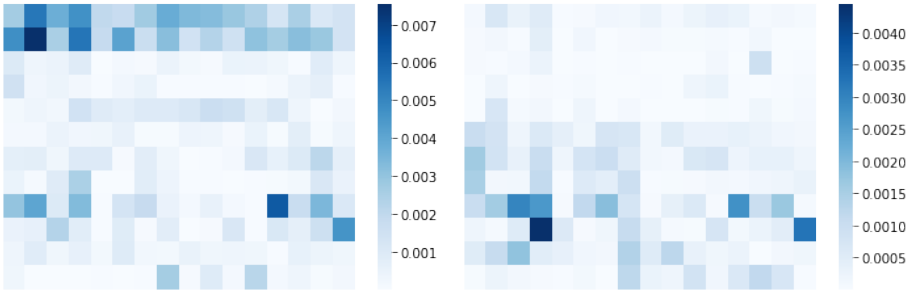
4. ábra. Az eredeti spektrogramtól vett különbség 500 (balra) és 5000 (jobbra) epoch után, L1 normát használva, 40 pixelenként aggregálva



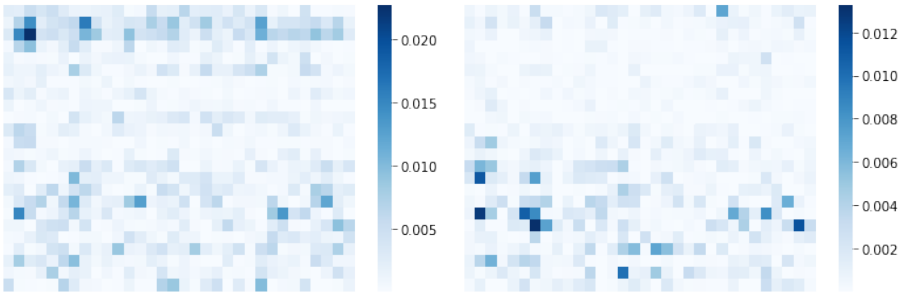
5. ábra. Az eredeti spektrogramtól vett különbség 500 (balra) és 5000 (jobbra) epoch után, L1 normát használva, 20 pixelenként aggregálva



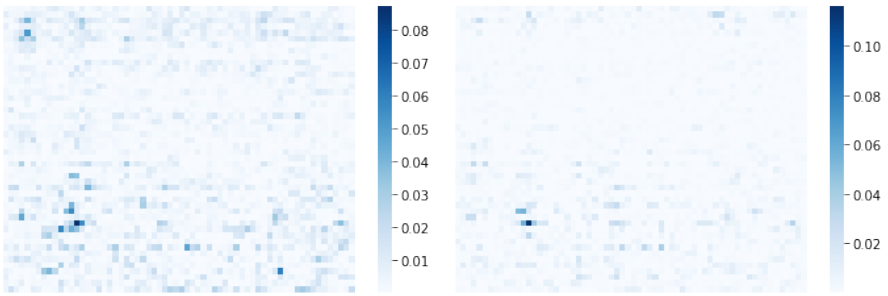
6. ábra. Az eredeti spektrogramtól vett különbség 500 (balra) és 5000 (jobbra) epoch után, L1 normát használva, 10 pixelenként aggregálva



7. ábra. Az eredeti spektrogramtól vett különbség 500 (balra) és 5000 (jobbra) epoch után, L2 normát használva, 40 pixelenként aggregálva



8. ábra. Az eredeti spektrogramtól vett különbség 500 (balra) és 5000 (jobbra) epoch után, L2 normát használva, 20 pixelenként aggregálva



9. ábra. Az eredeti spektrogramtól vett különbség 500 (balra) és 5000 (jobbra) epoch után, L2 normát használva, 10 pixelenként aggregálva

Hivatkozások

[1] Jason Brownlee - 2019 - A Gentle Introduction to Generative Adversarial Networks (GANs)

`https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/`

[2] Joseph Rocca - 2019 - Understanding Generative Adversarial Networks (GANs)

`https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29`

[3] Ketan Doshi - 2021 - Audio Deep Learning Made Simple

`https://ketanhdoshi.github.io/Audio-Augment/`

[4] Jungil Kong, Jaehyeon Kim, Jaekyoung Bae - 2020 - HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis - 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

`https://arxiv.org/abs/2010.05646`

[5] Hannah Weller - 2021 - Color Distance Metrics

`https://cran.r-project.org/web/packages/colordistance/vignettes/color-metrics.html`