

**Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky**

Vizualizácia hudby

Bakalárska práca

2018

Marián Sabat

**Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky**

Vizualizácia hudby

Bakalárska práca

Študijný program: Informatika
Študijný odbor: 9.2.1. Informatika
Školiace pracovisko: Katedra počítačov a informatiky (KPI)
Školiteľ: Ing. Norbert Ádám, PhD.
Konzultant:

Košice 2018

Marián Sabat

Názov práce: Vizualizácia hudby

Pracovisko: Katedra počítačov a informatiky, Technická univerzita v Košiciach

Autor: Marián Sabat

Školiteľ: Ing. Norbert Ádám, PhD.

Konzultant:

Dátum: 25. 5. 2018

Kľúčové slová:

Abstrakt: ABSTRAKT

Thesis title: Music Visualizer

Department: Department of Computers and Informatics, Technical University of Košice

Author: Marián Sabat

Supervisor: Ing. Norbert Ádám, PhD.

Tutor:

Date: 25. 5. 2018

Keywords:

Abstract: ABSTRAKT

Tu vložte zadávací list pomocí příkazu
`\thesispec{cesta/k/suboru/so/zadavacim.listom}`
v preambule dokumentu.

Čestné vyhlásenie

Vyhlasujem, že som záverečnú prácu vypracoval(a) samostatne s použitím uvedenej odbornej literatúry.

Košice, 25.5.2018

.....

Vlastnoručný podpis

Podakovanie

Obsah

Úvod	1
1 Formulácia úlohy	3
2 Syntéza dát	4
2.1 Zmena textu na reč	4
2.2 Prevod reči na text	5
2.3 Rozpoznávanie obsahu dát	5
2.4 Generovanie obrázkov	6
3 Teoretický základ	7
3.1 Hudobné dáta	7
3.2 Neurónové siete	8
3.3 Variačné autoenkódery	10
3.4 Generatívne konkurenčné siete	11
4 Tvorba datasetu	12
4.1 Zvukové dáta	12
4.2 Obrazové dáta	13
4.3 Zber dát	14
5 Detekcia farby	16
5.1 Návrh	16
5.2 Implementácia	17
Literatúra	19

Zoznam obrázkov

3.1	Model formálneho neurónu.	9
5.1	Graf aktivačnej funkcie ReLu.	17

Zoznam tabuliek

3.1	Akustické vlastnosti zvukových signálov	8
-----	---	---

Úvod

Systémy, ktoré sa dokážu učiť z dát sú už dnes prístupné verejnosti. Je čoraz jednoduchšie študovať techniky strojového učenia a preto progres v tomto odvetví je skutočne viditeľný. Množstvo dát a dobrá výpočtová technika majú za následok, že v takmer všetkých oblastiach sa zavádza nejaký druh umelej inteligencie. Počítače, ktoré by rozumeli informáciám by znamenali revolúciu v našich životoch. Program, ktorý by dokázal vygenerovať obraz na základe hudobného podkladu, na základe emócií a nálad, ktoré sú obsiahnuté v hudbe by bol pokrok ku umelej inteligencii, ktorá by skutočne rozumela dátam.

Proces syntézy jedného druhu informácií na iní je pre ľudí prirodzený no pre stroj je to neľahká úloha. Avšak progres v neurónových sieťach a v generatívnych algoritmoch umožňuje klasifikáciu jednej informácie a jej následnú zmenu na inú formu. Stále ale existuje množstvo prekážok v realizácii tohto problému.

To všetko nás privádza k otázke, sú dnešné neurónové siete schopné previesť hudobnú skladbu na zmysluplný obraz? Prevod hudobnej informácie na obrazovú si vyžaduje určitý stupeň kreativity a znalostí. V našej práci sa budeme snažiť zodpovedať tento problém. Budeme sa snažiť vytvoriť model, ktorý by simuloval ľudskú kreativitu.

V prvom rade je dôležité upraviť dáta, ktoré budeme analyzovať. Ide o zvukové signály, ktoré ako také sú nespracovateľné dnešnými algoritmami strojového učenia. Je nevyhnutné aby sme tieto dáta upravili na použiteľnú formu. Ďalším krokom je zistenie či sú počítače vôbec schopné priradenia najjednoduchšej obrazovej formy, čiže farby, k hudobným skladbám. Úspešné splnenie tejto úlohy bude dobrým predpokladom pre vytvorenie finálneho modelu, ktorý dokáže generovať obrázky na vyššej kreatívnej úrovni.

Naša práca je preto rozdelená presne podľa týchto celkov. Prvé kapitoly poskytnú súčasné úspechy v syntéze dát a teoretický základ pre naše riešenie. V

ďalších kapitolách postupne prejdeme naše výsledky od najjednoduchších modelov až po tie zložitejšie. Na konci poskytneme porovnanie nami vytvorených systémov a odvodenie záverov.

1 Formulácia úlohy

2 Syntéza dát

Už od čias pred počítačmi ľudia využívali abstrakciu skutočných dát v podobe čísel a matematiky. S vývojom výpočtovej techniky prišli aj nové spôsoby zmeny jedného typu informácií na iný. Dnes existuje mnoho systémov určených na tento proces.

2.1 Zmena textu na reč

Už v osemdesiatych rokoch dvadsiateho storočia, keď Steve Jobs predstavil nový Macintosh, počítač vedel hovoriť. Systémy, ktoré používajú zmenu textu na reč sú dnes bežná vec, a preto je ťažké predstaviť si svet bez nich [1]. Aj keď sú zaužívané stále existuje priestor na zlepšenie. Hlas starého Macintosha bol zreteľne umelý, no dnes existujú programy, ktoré dokážu simulovať ľudskú reč takmer na nerozpoznanie od živých ľudí. Tieto syntetizátory našli svoje využitie napríklad v telekomunikáciách. Primitívne úlohy vykonávané cez telefón sú zverené počítačom. Ľudia si už zvykli, že keď volajú niekam aby si niečo vybavili je normálne ak sa im ozve stroj. Syntetizátory našli svoje využitie aj vo vzdelávaní. Učenie jazykov z pohodlia domova je možné aj vďaka tomu, že počuť ako sa slovo vyslovuje môžeme bez prítomnosti skúseného rečníka či cudzinca. Zrakovo postihnutí práve vďaka technológiám zmeny textu na reč môžu využívať prístroje, ako napríklad počítače, telefóny a iné, bez ktorých sa dnes už nezaobídeme. Nie len slepým, ale aj nemým a inak telesne postihnutým pomáhajú čítačky textu každý deň. Svoje využitie našli v mnohých oblastiach od vedy a výskumu až po zábavný priemysel.

2.2 Prevod reči na text

Pre ľudí veľmi jednoduchá úloha, rozpoznanie reči, je pre číselné systémy netriviálna záležitosť. Fourierové transformácie a úprava dát nám dávajú šancu na extrakciu vhodných informácií, ktoré sa stali vhodným nástrojom na detekciu slov [2]. Systém schopný prevodu hovorenej reči na text je nevyhnutnosťou v prípadoch, keď človek nemôže alebo nedokáže použiť klávesnicu či iní mechanický vstup. Osobní asistenti ako Cortana alebo Ok Google by bez týchto syntetizátorov nedosiahli takej popularity. Prevodníky reči na text majú veľký vplyv na to ako pracujeme s našimi zariadeniami. Funkcie ako preklad z jedného jazyka do druhého v reálnom čase sa zavádzajú do programov určených na komunikáciu a znižujú tak priepasť medzi ľuďmi rôznych národností. To by nebolo možné ak by jadrá týchto programov nestáli na technológiách prevodu reči na text a textu na reč.

2.3 Rozpoznávanie obsahu dát

S úlohou syntézy informácií súvisí aj problém reprezentovania významu dát. Naučiť počítače rozpoznávať čo sa nachádza na obrázku je dnes silno skúmaná oblasť. Mnohé automobilové spoločnosti sa snažia vytvoriť samo jazdiace vozidlá, ktoré dokážu spozorovať prekážky okolo seba a zareagovať rýchlejšie ako by to dokázal ktorýkoľvek človek. Vďaka najnovším metodikám, ako napríklad využite konvolučných neurónových sietí, ľudia vytvorili programy, ktoré dokážu rozpoznať čo sa nachádza na obrázkoch a rozoznať jednotlivé objekty [3]. Takéto technológie sa dajú využiť napríklad pri vyhľadávaní. Spoločnosti ako Google ponúkajú cloudové služby stvorené presne na tieto účely [4].

Neurónové siete dokážu analyzovať obrázky na úrovniach aké doposiaľ neboli možné. V roku 2016 výskumníci z Montrealu a Toronta vytvorili model, ktorý dokázal analyzovať obrázky a vytvoriť textový popis týkajúci sa obsahu obrázku [5]. V spojení s čítačkou textu by sme takto mohli ešte viac uľahčiť prístup k informáciám aj pre zdravotne postihnutých ale aj využiť takúto technológiu na ešte lepšie výsledky. Detekcia tváre sa zavádza aj do mobilných telefónov a využíva sa na odomknutie uzamknutej obrazovky. Facebook vytvára systém, ktorý by využil rozpoznanie tváre a využil tieto dáta ako heslo pre používateľa [6]. Analýza

obrazu ale neostáva len pri rozpoznávaní objektov. Inteligentné systémy dokážu rozpoznať rôzne vlastnosti obrazu. Google vytvoril systém, ktorý nazval Deep dream. Deep dream dokáže rozpoznať vlastnosti obrazu a upraviť pôvodný obrázok pridaním vrstiev, ktoré majú podobné vlastnosti. Vytvorené obrázky potom vyzerajú ako halucinácie. Mnohé iné webové aplikácie zase využívajú siete, ktoré sa naučia ako rozpoznať štýl obrazu a vďaka tomu dokážu preniesť štýl na úplne iný obrázok. Vznikajú tak zaujímavé filtre na úpravu obrázkov a fotiek.

2.4 Generovanie obrázkov

V posledných rokoch systémy strojového učenia preukazujú výsledky v generovaní nových dát. Syntéza textu na obraz je jednou z najnovších prác v tomto odvetví [7]. Nakoľko ide o experiment, tak komerčné využitie ešte neexistuje. Ale pri zlepšení tejto technológie sa môže vytvoriť obrovský potenciál. Takéto výskumy prebiehajú na celom svete a na prechod na trh určite nebudeme dlho čakať. Umelo generované obrázky dokážeme vďaka tomu ako sú vytvorené ďalej upraviť. Nakoľko sú to obrázky vytvorené pomocou matematického modelu vieme upravovať obsah veľmi jednoducho. Príkladom sú generované obrázky ľudí, v ktorých sa dá upravovať napríklad to či sa osoba usmieva alebo nie, či má okuliare alebo mnoho iných vlastností [8].

3 Teoretický základ

Ako už bolo avizované naša práca sa zaoberá neurónovými sieťami. Preto v tejto kapitole zhrnieme základné pojmy a problémy, pred ktorými stojíme.

Algoritmus, ktorý sa učí z dát, poskytuje dobré výsledky len vtedy ak má dobré dáta, z ktorých sa učí. Naš model sa bude učiť z hudobných dát. Zvukové analógové signály uložené v digitálnej forme sú príliš rozsiahle nato aby sme ich mohli využiť ako vstup do neurónovej siete. Jedinou možnosťou je úprava takýchto dát na jednoduchšiu formu.

3.1 Hudobné dáta

Ľudia vnímajú hudbu na rôznych úrovniach. Rôzne vlastnosti harmonických zvukov vyvolávajú u nás rozličné emócie. Napríklad kým vzrušenie blízko súvisí s tempom, tak výška tónov a hlasitosť skôr určujú náladu skladby [9]. To ako vnímame hudbu má preto veľký vplyv na našu predstavivosť.

Hudobné dáta sú ľahko dostupné na internete. Čo sa týka hudby existuje niekoľko zdrojov pre informácie [10]:

- Hudobné metadáta,
- Akustické vlastnosti,
- Slová,
- Hudobná kritika,
- MIDI súbory,
- Hudobné skóre.

Tabuľka 3.1: Akustické vlastnosti zvukových signálov

Sada vlastností	Extrahovateľné vlastnosti
Energia	Dynamická hlasitosť, výkon zvuku, celková hlasitosť, špecifické koeficienty citlivosti na hlasitosť
Rythmus	Diagram úderov, vzor rytmu, histogram rytmu a tempo, sila rytmu, pravidelnosť rytmu, jasnosť rytmu, priemerná frekvencia nástupu, priemerné tempo
Časové vlastnosti	Nulové priechody, logaritmus času útoku
Spektrálne vlastnosti	Spektrálne centroidy, spektrálne vyhodnocovanie, spektrálny tok, spektrálne merania rovinnosti, spektrálne hrudkové faktory, mel-frekvenčné spektrálne koeficienty
Harmónia	Jasnosť klúča, hudobný režim, harmonická zmena, diagram stúpania

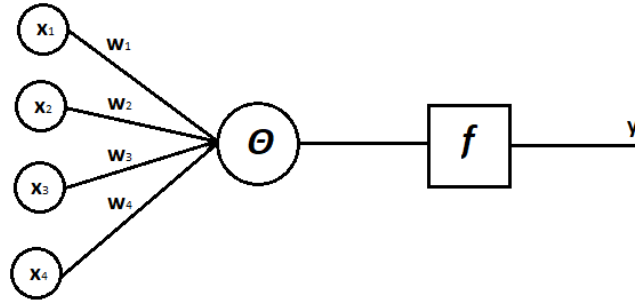
Rôzne typy informácií majú využitie v rôznych prípadoch. Nás budú zaujímať akustické vlastnosti, medzi ktoré patrí energia, rytmus, časové a spektrálne vlastnosti [9]. Pre každú z týchto sád vlastností existuje niekoľko vlastností, ktoré sa dajú extrahovať známymi algoritmami a programami. V tabuľke 3.1 sme uviedli zopár významných údajov, ktoré by sme mohli využiť pre tréning našich modelov. Tieto údaje opisujú hudobnú informáciu a majú značne menšiu veľkosť ako skladba uložená vo formáte mp3 alebo wav.

3.2 Neurónové siete

Tak ako mozog, neurónová sieť je spojenie viacerých neurónov do siete. Neurón na vstupe prijíma informácie, ktoré pozmení a pošle na výstup, čo môže byť vstup ďalšieho neurónu.

Formálny neurón alebo perceptrón (obrázok 3.1) má na vstupe n aktivít [11]. Označme vstup ako vektor $x = (x_1, x_2, x_3, \dots, x_n)^T$, kde T označuje transponovaný vektor.

Všetky vstupné kanály majú svoju váhu, označme preto váhy vstupov ako vektor $w = (w_1, w_2, w_3, \dots, w_n)^T$. Celkový vstup perceptrónu sa potom vypočíta ako sú-



Obr. 3.1: Model formálneho neurónu.

čet súčinu vektorov x a w , a prahu excitácie Θ . Pre získanie výstupu vstupy prejdú aktivačnou funkciou. Pre výstup y potom platí

$$y = f(w^T x + \Theta)$$

Neuróny v sieťach sa spájajú do vrstiev. Nech je vrstva tvorená m neurónmi a každý neurón má n vstupných kanálov. Označme váhu j -teho vstupného kanála do i -teho neurónu ako w_{ij} . Potom môžeme vytvoriť váhovú maticu

$$W = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & & & \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix}$$

Výstup takejto vrstvy potom bude $y = (y_1, y_2, y_3, \dots, y_m)^T$, ktorý dostaneme ako

$$y = Wx$$

Tento výstup potom slúži ako vstup do ďalšej vrstvy, ktorej celkový počet vstupných kanálov je rovný m .

Po prechode celou sieťou sa vypočíta strata s ohľadom na požadovaný výstup. Trénovanie takéhoto modelu je minimalizačný problém kedy sa upravujú hodnoty váhových matic a prahov aktivácie tak aby bola strata čo najmenšia.

Neurónové siete sú výborné klasifikátory ale dajú sa použiť aj na generovanie nových dát. Existuje niekoľko druhov sietí, ktoré dokážu generovať obrázky. Pre nás zaujímavé budú VAE, CPPN a GAN neurónové siete.

3.3 Variačné autoenkóдеры

VAE alebo Variačné autoenkóдеры (angl. Variational autoencoders) vznikli upravením jednoduchých autoenkóderov [12]. Neurónová sieť postavená ako autoenkóder sa dokáže naučiť štruktúru vstupných dát. Pre jednoduché vysvetlenie majme na vstupe do siete rastrový obrázok. Obrazové dáta sú uložené vo forme pixlov, ktorých množstvo závisí od rozmerov obrázka. Autoenkóder dokáže uložiť štruktúru umiestnenia pixlov v obrázku, čiže jeho obsah do jednoduchých premenných nazvaných latentné premenné. Vektor tvorený z týchto premenných je komprimáciou obrázka a z tohto vektora môžeme dekódovaním dostať pôvodný obrázok. Ak by sme vedeli ako jednotlivé dimenzie latentného vektora ovplyvňujú obrázok, mohli by sme meniť jeho vlastnosti jednoduchou zmenou premennej. Tento fakt sa dá využiť pri generovaní nových dát.

V prvom rade pre všetky naše dáta X v datasete musí existovať nastavenie latentných premenných, ktoré umožňuje modelu generovať niečo veľmi podobné naším dátam. Majme vektor latentných premenných z v multidimenzionálnom priestore Z , ktorý môžeme ľahko vybrať podľa nejakej funkcie hustoty pravdepodobnosti $P(z)$ definovanej nad Z . Majme funkciu $f(z, \theta)$ parametrizovanú vektorom θ v priestore Θ , kde $f : Z \times \Theta \rightarrow X$. Ak je z náhodné a θ nemenné, potom $f(z, \theta)$ je náhodná premenná v priestore X . My chceme optimalizovať θ tak, že ak vyberieme vzorku z $P(z)$, tak $f(z, \theta)$ bude podobné naším dátam X .

VAE sieť musí zistiť akú informáciu nesie latentná premenná. VAE majú neobvyklý prístup k tejto úlohe. Pri variačných autoenkódroch predpokladáme, že neexistuje jednoduchá interpretácia dimenzií z . Namiesto toho sa vzorky z berú z Gaussovho normálneho rozdelenia.

V praxi to vyzerá, tak že pri kódovaní dát na vektor z sa vytvárajú dva vektory, vektor stredných hodnôt a vektor smerodajnej odchýlky. Tieto vektory potom spoločne tvoria výsledný vektor z . Podľa vektora stredných hodnôt sa vypočítava strata, ktorá hovorí či sú vygenerované dáta podobné chceným dátam. A podľa vektoru smerodajnej odchýlky sa vypočíta strata, ktorá meria ako blízko sú latentné premenné k normálnemu rozdeleniu. Celková strata je suma týchto čiastkových strát.

Nevýhodou takýchto sietí je že ak sa použijú na generovanie obrázkov, tak generované obrázky sú rozmazané práve kvôli strate pri kompresii, ktorú VAE

vytvárajú.

3.4 Generatívne konkurenčné siete

Generačné konkurenčné siete alebo GAN (angl. Generative adversarial networks) pôvodne navrhnuté Ianom Goodfellowom, fungujú na princípe konkurencie medzi dvoma sieťami [13]. Prvá sieť, generátor, sa snaží generovať čo najrealistickejšie dáta zo šumu. Druhá sieť, diskriminátor, rozpoznáva či sú vstupy reálne alebo vygenerované. Tieto siete sa tak snažia poraziť jedna druhú. GAN sú známe generovaním takmer realistických obrázkov. No nevýhodou je, že sa veľmi ťažko trénujú. Je viacero situácií, ktoré môžu nastať. Môže sa stať, že generátor nájde systém ako oklamať diskriminátor a generuje len jeden druh informácií. Alebo diskriminátor sa stane tak dobrým v rozpoznávaní skutočnosti, že vyhodnotí všetky generované dáta za neskutočné a tak sa generátor nebude môcť zlepšiť.

GAN zaznamenali niekoľko vylepšení a využití v rôznych oblastiach. Pri generovaní obrázkov sa osvedčilo využitie konvolučných vrstiev [8]. Takéto siete dokážu generovať oveľa kvalitnejšie obrázky aj keď stále nie vo veľkom rozlíšení. Ďalším vylepšením, práve v tomto probléme bolo využitie viacerých GAN sietí [14]. Zreťazenie viacerých neurónových sietí má za následok lepšiu kvalitu generovaných obrázkov. Niekoľko vrstiev dokáže zväčšiť obrázok s rozmermi 64x64 na rozmery 256x256 a výrazne upraviť kvalitu. A v poslednom rade je aj výskum podmienených GAN. Tie vznikajú pridaním vlastností na vstup, pričom sa takto dá viac, či menej ovplyvniť výstup [15]. Práca z Univerzity v Michigane využíva práve podmienené GAN. V tejto práci využili neurónovú sieť na syntézu textu na obraz. Z datasetu vtáctva a kvetín vytvorili model, ktorý dokáže generovať obraz podľa popisu [7]. S dostatočne veľkými zdrojmi by mohol mať tento prístup veľké využitie.

4 Tvorba datasetu

Najdôležitejšou časťou pri vytváraní nášho systému je tvorba datasetu. Podľa toho aké dobré dáta zozbierame, také výsledky sa nám podarí získať. Náš dataset bude tvorený dvomi druhmi dát, a to obrazovými a zvukovými.

4.1 Zvukové dáta

V predošlej kapitole sme vysvetlili prečo a ako je potrebné upraviť hudobné dáta, z ktorých sa bude náš model učiť. V tejto podkapitole opíšeme ako bude vyzeráť zvuková časť nášho datasetu.

Prvým krokom je výber žánru skladieb, ktoré budeme zbierať. V súčasnosti existuje nesmierne množstvo interpretov a skladieb z ktorých si môžeme vybrať. No nie každý hudobný žáner dokáže podnietiť našu predstavivosť. Pre náš projekt budú najvhodnejšie skladby klasickej hudby, nakoľko na rozdiel od populárnej modernej hudby, klasická hudba dáva väčší dôraz na emócie.

Vzorky datasetu by mali mať konštantnú veľkosť. Preto je nevyhnutné upraviť skladby tak aby tomu vyhovovali.

V závislosti od dĺžky a rôznorodosti jednotlivých skladieb vyberieme z každej niekoľko vzoriek s dĺžkou 25 sekúnd. Táto dĺžka je zvolená preto aby mal poslucháč dostatok času na vybavenie obrazu. Pomocou programu Audacity ďalej upravíme každú vzorku pomocou efektu Normalizovať. Takto dosiahneme podobnú hlasitosť, bez ohľadu na spôsob nahrávania daných vzoriek. Taktiež zmeníme vzorkovaciu frekvenciu na 22 050Hz, aby naša analýza bola súvislá.

Pre hlavnú analýzu vytvoríme skript v programovacom jazyku Python. Tento jazyk sme vybrali kvôli možnosti modulov, ktoré existujú a pomôžu nám pri tvorbe systému. Pre hudobnú analýzu použijeme modul Librosa, ktorý obsahuje množstvo algoritmov vhodných pre získanie vlastností akustických signálov, ktoré sme

uviedli v predošlej kapitole.

Vybrali sme päť akustických vlastností, ktoré budú reprezentovať naše skladby, a to mel-frekvenčné kepstrálne koeficienty, tempo, spektrálne centroidy, počet nulových priechodov a mieru týchto priechodov. Každá z týchto vlastností nesie užitočnú informáciu potrebnú pre našu úlohu. Napríklad mel-frekvenčné kepstrálne koeficienty poskytujú informáciu o tom aký tón a farbu má skladba v danom časovom úseku.

Touto analýzou sme úspešne skrátili vstupný vektor. Ak by sme chceli vložiť 25 sekundové vzorky skladieb do neurónovej siete celkový počet vstupných kanálov by bol $25 \times 22050 = 551250$. Náš dataset obsahuje pre každú vzorku 549 hodnôt z oboru reálnych čísel. Mel-frekvenčné kepstrálne koeficienty tvoria najväčšiu časť.

```
S = librosa.feature.melspectrogram(
    y=y, sr=sr, n_mels=128, hop_length=22050, n_fft=22050)
log_S = librosa.logamplitude(S, ref_power=np.max)
mfcc = librosa.feature.mfcc(S=log_S, sr=sr, n_mfcc=20)
mfcc_1d_vector = mfcc.flatten()
```

Tie sú vypočítané na základe mel-spektrogramu so vzorkovaním 22050, čo vytvára 26 hodnôt. Počet mfcc je 20, celkovo tak dostávame $20 \times 26 = 520$ čísel. Tempo, aritmetický priemer spektrálnych centroidov a celkový počet nulových priechodov sú ďalšie 3 hodnoty datasetu. Posledných 26 hodnôt tvoria miery nulových prechodov, merané so skokom 22050.

```
zcr = librosa.feature.zero_crossing_rate(
    y, frame_length=22050, hop_length=22050).flatten()
```

4.2 Obrazové dáta

Tak ako hudobné dáta tak aj obrázky, ktoré bude náš dataset obsahovať musia mať nejakú koherentnú formu.

Prvým modelom, ktorý budeme vytvárať bude neurónová sieť, ktorá dokáže priradiť k skladbe farbu. Pre tento klasifikátor ako výstup použijeme jedno číslo, ktoré bude reprezentovať farbu v našej farebnej palete. Takto sme sa rozhodli z dôvodu, že pri vytváraní datasetu bude jednoduchšie ak k jednotlivým skladbám priradíme farbu z vopred danej palety.

Pre druhý a náš hlavný model, budeme ukladať obrázky vo formáte png a každý obrázok bude uložený vo veľkosti 32x32 pixlov.

4.3 Zber dát

V tejto časti opíšeme spôsob ako sme zbierali dáta pre náš dataset. Zber údajov je veľmi podstatná časť projektu. Preto je nevyhnutné nebrať tento proces na ľahkú váhu. Vytváranie vlastností (features) a štítkov (labels) môže byť úplne manuálne alebo z časti automatizované. Zber dát manuálne môže byť veľmi pracný a zdĺhavý, preto je vhodné mať nejaký nástroj pre túto činnosť. Tento nástroj by nás oslobodil od množstva repetitívnej práce. Ďalšou výhodou pre vytváranie datasetu na jednom mieste je zapojenie viacerých ľudí do procesu. Viac ľudí prináša variabilitu do našich dát a umožňuje vytvorenie robustnejšieho riešenia.

Z týchto dôvodov sme vytvorili webovú aplikáciu, ktorá slúži na zber údajov. V našom datasete sú okrem akustických vlastností skladieb, ktoré dokážeme získať relatívne jednoducho, aj údaje, ktoré nedokážeme získať pomocou algoritmov. Presne pre získanie týchto dát sme vytvorili formulár určený na ich ukladanie. Konkrétne ide o farbu skladby, emóciu a obrázok, toho čo skladba predstavuje. Farbu je možné vybrať z palety pätnástich základných farieb. Pre emóciu možno vybrať jednu z piatich skupín emócií. Jednotlivé skupiny možno opísať ako vášnivé emócie, veselé, dojímavé, vtipné a agresívne. Najdôležitejšou informáciou, ktorú potrebujeme získať je ale obrázok. Do nášho datasetu potrebujeme dva druhy obrázkov, a to jednoduchú skicu a detailný obrázok. Pre jednoduché stvárnenie skladby v podobe skice sme do formuláru pridali editor na nakreslenie obrázku. Kvôli jednoduchosti je možné požiť len čiernu farbu. V datasete potrebujeme obrázky o veľkosti 32x32, lenže kreslenie takýchto obrázkov by bolo nepraktické. Preto sme plátnu na kreslenie dali veľkosť 320x320 a zväčšili sme hrúbku štetca. Takto budeme môcť kresliť pohodlne a zároveň potom zmenšiť obrázky bez prílišnej straty. Druhý typ obrázku získavame pomocou url adresy. Detailnejšie obrázky vyhladáme na internete a pomocou nášho nástroja uložíme adresu. Databáza takto nebude musieť ukladať surové dáta a my budeme môcť obrázok jednoducho vyhladať a stiahnuť. Keďže naša aplikácia verejne dostupná, pribudla nám nová úloha, a to kontrola. Skôr ako budeme môcť využiť nazbierané dáta, musíme ich skontrolovať či neobsahujú nevhodný obsah a či sú vierohodné.

Pomocou tejto webovej aplikácie sme zozbierali potrebné údaje, ktoré je nutné už len stiahnuť a jednoducho upraviť pre vstup do modelov neurónových sietí.

5 Detekcia farby

V predchádzajúcej kapitole sme opísali ako vyzerá náš dataset ako sme ho vytvorili. Táto kapitola je venovaná nášmu prvému modelu neurónovej siete. Ešte pred vytvorením zložitých a výpočtovo náročných algoritmov, určených na vizualizáciu hudby, bolo nevyhnutné dokázať, že naše dáta nesú informáciu, ktorá dovoľuje takýto počin.

Najjednoduchšia forma vizualizácie hudobnej skladby je reprezentácia farbou. Pri počúvaní hudby si ľudský mozog dokáže predstaviť obraz. Tento obraz nemusí obsahovať konkrétne objekty, ale môže byť čisto abstraktný s veľkou dominanciou farby. Náš algoritmus by mal dokázať podobnú vec.

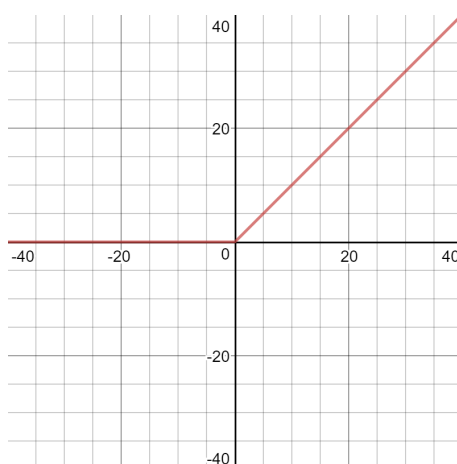
5.1 Návrh

Na klasifikáciu sme využili doprednú hlbokú neurónovú sieť s tromi skrytými vrstvami a jednou výstupnou vrstvou. Pričom sa každá skrytá vrstva skladá zo sto neurónov. Na vstupe očakávame dáta na 549 kanáloch. Výstupom je vektor y s päťnástimi dimenziami. Veľkosť tohto vektora súvisí s využitím techniky One-Hot-Encoding. Zo vstupných dát model určuje jednu z päťnástich farieb, preto je pre každú farbu priradený jeden rozmer. Aktivačnou funkciou pre jednotlivé vrstvy je ReLu (obrázok 5.1). ReLu je definovaná ako

$$g(x) = \max(0, x)$$

Pričom obor hodnôt je $H =]-\infty, \infty[$.

Trénovanie modelu je realizované v troch krokoch. Výpočty sú vykonávané na dávkach dát. V prvom kroku je vytvorená predikcia výsledku. Následne sú tieto výsledky porovnané s očakávanými výstupmi a vypočíta sa chyba. Táto chyba je vstupom do optimalizátora, ktorý upraví premenné siete. Celý proces sa opakuje vo viacerých epochách pre dosiahnutie väčšej presnosti.



Obr. 5.1: Graf aktivačnej funkcie ReLu.

Na výpočet rozdielu medzi výstupom neurónovej siete a predpokladaným výstupom využijeme krížovú entropiu. Krížová entropia pre dve diskkrétne rozdelenia pravdepodobností p a q je definovaná ako

$$H(p, q) = - \sum p(x) \log q(x)$$

Optimalizovanie premenných modelu, čiže váhových matic a prahov excitácie zabezpečí algoritmus spätného šírenia chyby. Konkrétne ide o modifikáciu algoritmu Gradient Descent.

5.2 Implementácia

Model klasifikátora a jeho tréning je implementovaný v jazyku python, využitím modulu TensorFlow. TensorFlow obsahuje mnoho algoritmov strojového učenia, preto nás oslobodzuje od ich implementovania. Ďalšou výhodou tohto modulu je to, že separuje výpočty od pythonu, preto sú tieto výpočty rýchlejšie a je možné ich paralelizovať. Program pozostáva z dvoch krokov, vytvorenie výpočtového grafu a jeho spustenie. Výpočtový graf obsahuje celý model neurónovej siete spolu so všetkými výpočtovými krokmi pre tréning.

Vďaka tomu je možné implementovať funkciu pre tréning len niekoľkými riadkami.

```
with tf.Session() as sess:
    sess.run(tf.global_variables_initializer())
```

```
for epoch in range(hm_epochs):
    for _ in range(num_examples/batch_size):
        epoch_x, epoch_y = next_batch(batch_size)
        sess.run([optimizer, cost],
                  feed_dict={x: epoch_x, y: epoch_y})
```

Na začiatku je nutné nainicializovať premenné, až potom je možné s nimi pracovať. V každej epoche sú prejdené všetky dáta a sú optimalizované premenné siete. Výpočty sa spúšťajú príkazom `run`. Jeho parametrami sú výpočty, vrcholy TensorFlow grafu, a vstupné premenné potrebné pre tieto výpočty. Optimizer je verzia algoritmu Gradient Descent.

```
optimizer = tf.train.AdamOptimizer().minimize(cost)
```

Na výpočet chyby (`cost`) je využitá TensorFlow implementácia krížovej entropie.

```
prediction = neural_network_model(x)
cost = tf.reduce_mean(tf.nn.softmax_cross_entropy_with_logits(
    logits=prediction,
    labels=y))
```

Literatúra

- [1] Thierry Dutoit. *A Short Introduction to Text-to-Speech Synthesis*. Online. Dec. 1999. URL: http://tcts.fpms.ac.be/synthesis/introtts_old.html.
- [2] Fifth Generation Computer Corporation. *Speaker independent connected speech recognition*. Online. URL: <http://www.fifthgen.com/speaker-independent-connected-s-r.htm>.
- [3] *Image Recognition*. TensorFlow. Nov. 2017. URL: https://www.tensorflow.org/tutorials/image_recognition.
- [4] Google Cloud Platform. *Cloud vision API*. Online. URL: <https://cloud.google.com/vision/>.
- [5] Ryan Kiros Kelvin Xu Jimmy Lei Ba. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. Tech. spr. Université de Montréal, University of Toronto, apr. 2016. URL: <https://arxiv.org/pdf/1502.03044.pdf>.
- [6] Matthew Field. "Facebook tests face recognition technology". In: *The Telegraph* (okt. 2017). URL: <http://www.telegraph.co.uk/technology/2017/10/02/forgot-password-facebook-tests-face-recognition-technology-unlock/>.
- [7] Xinchun Yan Scott Reed Zeynep Akata. *Generative Adversarial Text to Image Synthesis*. Tech. spr. University of Michigan, Ann Arbor, MI, USA, jún 2016. URL: <https://arxiv.org/pdf/1605.05396.pdf>.
- [8] Soumith Chintala Alec Radford Luke Metz. *Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks*. Tech. spr. in-dico Research, Facebook AI Research, jan. 2016. URL: <https://arxiv.org/pdf/1511.06434.pdf>.

- [9] Homer H. Chen Yi-Hsan Yang. *Music Emotion Reckognition*. Boca Raton: CRC Press, 2011. ISBN: 978-1-4398-5046-6.
- [10] George Tzanetakis Tao Mitsunori Ogiwara. *Music data mining*. Boca Raton: CRC Press, 2012. ISBN: 978-1-4398-3552-4.
- [11] Jiří Pospíchal Vladimír Kvasnička Ľubica Beňušková. *Úvod do teórie neurónových sietí*. Iris, 1997. ISBN: 8088778301.
- [12] Carl Doersch. *Tutorial on Variational Autoencoders*. Tech. spr. Carnegie Mellon / UC Berkeley, aug. 2016. URL: <https://arxiv.org/pdf/1606.05908.pdf>.
- [13] Jean Pouget-Abadie Ian J. Goodfellow. *Generative Adversarial Nets*. Tech. spr. Departement d'informatique et de recherche opérationnellé Université de Montréal, jún 2014. URL: <https://arxiv.org/pdf/1406.2661.pdf>.
- [14] Hongsheng Li Han Zhang Tao Xu. *StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks*. Tech. spr. Rutgers University, Lehigh University, The Chinese University of Hong Kong, aug. 2017. URL: <https://arxiv.org/pdf/1612.03242.pdf>.
- [15] Simon Osindero Mehdi Mirza. *Conditional Generative Adversarial Nets*. Tech. spr. Departement d'informatique et de recherche opérationnellé Université de Montréal, Flickr / Yahoo Inc. San Francisco, nov. 2014. URL: <https://arxiv.org/pdf/1411.1784.pdf>.