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 pomocou príkazu
`\thesisspec{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

Poďakovanie

Obsah

Zoznam obrázkov

Zoznam tabuliek

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 realizácií 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 probém. Budeme sa snažiť vytvoriť model, ktorý by symuloval ľudskú kreativitu.

V prvom rade je dôležité upraviť dáta, ktoré budeme analyzovať. Ide o zvukové sygnály, ktoré ako také sú nespracovateľné dnešnými algoritmami strojového učenia. Je nevihnutné 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 splenenie 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 kapytoly poskytnú súčasné úspechy v syntéze dát a teoretický základ pre naše riešenie. V ďal-

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

1 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.

1.1 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 postihnutý práve vďaka technológií zmeny textu na reč môžu využívať prístroje, ako napríklad počítače, telefóny a iné, bez ktorých sa dnes už nezaobídeme. Nie len slepím, 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.

1.2 Prevod reči na text

Pre ľudí veľmi jednoduchá úloha, rozpoznanie reči, je pre číselné systémy netriviálna záležitosť. Rozpoznanie nám známych rečových úkazov zo zvukovej vlny je takmer nemožné. 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. Osobný 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 zmenšujú 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.

1.3 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 rýchlejšia ako by to dokázal ktorýkoľvek človek. Vďaka najnovším metodikám, ako napríklad využite kovolučných 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 tvári sa zavádza aj do mobilných telefónov a využíva sa na odomykanie uzamknutej obrazovky. Facebook vytvára systém, ktorý by vy-

užil 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.

1.4 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 toku 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].

2 Teoretický základ

Ako už bolo avizované naša práca sa zaoberá neurónovými sieťami. Preto v tejto kapytole zhrnieme základné pojmi a problémi, pred ktorými stojíme.

Algoritmus, ktorý sa učí z dát, poskituje dobé výsledky len vtedy ak má dobré dáta, z ktorých sa učí. Náš model sa bude učiť z hudobných dát. Zvukové analógové sygnály 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.

2.1 Hudobné dáta

Ľudia vnímajú hudbu na rôznych úrovňach. 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 pliv 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 2.1: Akustické vlastnosti zvukových sygnálov

Sada vlastností	**Extrahovateľné vlastnosti**
Energia	Dynamická hlasitosť, výkon zvuku, celková hlasitosť, špecifické koeficienty citlivosti na hlasitosť
Rytmus	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 kepstrálne koeficienty
Harmónia	Jasnosť kľúča, hudobný režim, harmonická zmena, diagram stúpania

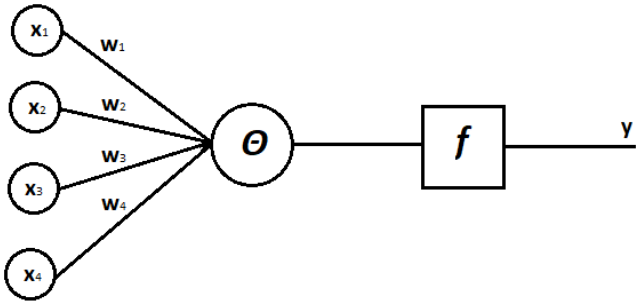
Rôzne typy informácií majú využitie v rôznych prípadoch. Nás budú zaujímať akustické vlastnosti, medzi ktoré patrí energia, rytmus, časové a spektrálne vlastnosti [9]. Pre každú z týchto sád vlastností existuje niekoľko vlastností, ktoré sa dajú extrahovať známimi algoritmami a programmi. V tabuľke 2.1 sme uviedli zopár významných údajov, ktoré by sme mohli využiť pre trénovanie našich modelov. Tieto údaje opisujú hudobnú informáciu a majú značne menšiu veľkosť ako skladba uložená vo formáte mp3 alebo wav.

2.2 Neurónové siete

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

Formálny neurón alebo perceptrón (obrázok 2.1) má na vstupe n aktivít [11]. Označme vstup ako vektor $x = (x_1, x_2, x_3, ...x_n)^T$, kde T označuje transponovaný vektor.

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



Obr. 2.1: 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 môžme 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 pomot bude $y = (y_1, y_2, y_3, ...y_m)^T$, ktorý dostanem ako

$$y = Wx$$

Tento výstup potom slúži ako vstup do dalšej 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 potom minimalizačný problém kedy sa upravujú hodnoty váhových matíc 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.

2.3 Variačné autoenkódery

VAE alebo Variačné autoenkódery (variational autoencoders) vznikli upravením jednoduchých autoencóderov [12]. Neurónová sieť postavená ako autoenkóder sa dokáže naučiť štruktúru vstupných dát. Pre jednoduché vysvetlenie majme na vstupe do siete obrázok. Obrazové dáta majú obrovské množstvo dimenzií vo forme pixelov. Autoenkóder dokáže uložiť štruktúru postavenia pixelov v obrázku 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 ovplivňujú obrázok, mohli by sme meniť jeho vlastnosti jednoduchou zmenou premennej. Tento fakt sa dá využiť pri generovaný 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 môžme 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 pemenná 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 neobviklí 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 vekory 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 kompresií, ktorú VAEs vytvárajú.

2.4 Generative Adversarial Networks

GANs alebo 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. GANs 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.

GANs 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]. Nakopenie 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 neposlednom rade je aj výskum podmienených GANs. Tie vznikajú pridaním vlastností na vstup, pričom sa takto dá viac, či menej ovplyvniť výstup [15]. Práca z Univerzity v Michigane využíva práve podmienené GANs. 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 a mohol by zmeniť mnoho systémov.

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 Comporation. *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. Universite 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] Xinchen 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. indico 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 Ogihara. *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 Universite 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 Universite de Montréal, Flickr / Yahoo Inc. San Francisco, nov. 2014. URL: https://arxiv.org/pdf/1411.1784.pdf.